NRL Report 8857

Low-Bit Rate Speech Encoders Based on Line-Spectrum Frequencies (LSFs)

G.S. KANG AND L.J. FRANSEN

Information Technology Division

January 24, 1985

NAVAL RESEARCH LABORATORY
Washington, D.C.

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	Approved for public release; distribution unlimited.
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NRL Report 8857	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Naval Research Laboratory 6b. OFFICE SYMBOL *(If applicable)*	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS *(City, State, and ZIP Code)* Washington, DC 20375-5000	7b. ADDRESS *(City, State, and ZIP Code)*
8a. NAME OF FUNDING/SPONSORING ORGANIZATION (See page ii) 8b. OFFICE SYMBOL *(If applicable)* PDE 110	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS *(City, State, and ZIP Code)* (See page ii)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO.: (see page ii) PROJECT NO.: (see page ii) TASK NO.: WORK UNIT ACCESSION NO.: DN 280-209

11. TITLE *(Include Security Classification)*
Low-Bit Rate Speech Encoders Based on Line-Spectrum Frequencies (LSFs)

12. PERSONAL AUTHOR(S)
Kang, G.S. and Fransen, L.J.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT *(Year, Month, Day)*	15. PAGE COUNT
Interim	FROM _____ TO _____	1985 January 24	63

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*	
FIELD	GROUP	SUB-GROUP	Line-spectrum frequencies	Vector quantization
			Speech analysis/synthesis	Linear predictive coding
			Low-bit rate speech encoder	Prediction residual coding

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

After nearly a dozen years of research and development, the narrowband linear predictive coder (LPC) operating at 2400 bits per second (b/s) has become a practical means to transmit speech at less than 5% of the bit rate of the original digitized speech. The 2400-b/s LPC is expected to be deployed extensively on various military platforms. Recently, however, there has been a growing interest in very-low-data-rate (VLDR) (800 b/s or less) voice communication for certain specialized military communications. Likewise, there is a demand for a voice processor operating at 4800 b/s that outperforms significantly the current 2400-b/s LPC for operation in less than favorable environments.

We present in this report a means for implementing 800- and 4800-b/s voice processors. Both processors use a similar speech synthesis filter in which control parameters are line-spectrum frequencies (LSFs). The LSFs are frequency-domain parameters transformed directly from the prediction coefficients
(Continues)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION	
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED	
22a. NAME OF RESPONSIBLE INDIVIDUAL Lawrence J. Fransen	22b. TELEPHONE *(Include Area Code)* (202) 767-2400	22c. OFFICE SYMBOL Code 7526

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted.
All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

8a. Name of Funding/Sponsoring Organization
Office of Naval Research
Naval Electronic Systems Command

8c. Address (City, State, and Zip Code)
Arlington, VA 22217
Washington, DC 20360

10. Source of Funding Numbers

Program Element No.
61153N
33904N

Project No.
RR021-05-42
X7290-CC

19. Abstract (Continues)

used in the conventional-narrowband LPC. The use of frequency-domain parameters is highly significant because they allow filter coefficient quantization in accordance with properties of auditory perception (i.e., coarser representation of higher frequency components of the speech spectrum). The excitation signal at the 800-b/s rate is similar to that used in the conventional-narrowband LPC. On the other hand, the excitation signal at 4800 b/s is the baseband residual signal.

At 800 b/s, speech intelligibility as measured by the diagnostic rhyme test (DRT) is 87 for three male speakers; this is about 1.4 points lower than that of the 2400-b/s LPC. At 4800 b/s, the DRT score is 92.3 which is 3.9 points higher than the score achieved by the 2400-b/s LPC and 0.7 of a point lower than the score achieved by the 9600-b/s Navy multirate processor. Both the 4800- and 9600-b/s voice processors have no voicing or pitch errors.

CONTENTS

LOW-BIT RATE SPEECH ENCODERS BASED ON LINE-SPECTRUM FREQUENCIES (LSFs)

INTRODUCTION

Voice data in military communications are increasingly being encoded by digital rather than analog waveforms because digital encryption for security reasons is both easier and less vulnerable to unauthorized decryption. The data rate of unprocessed speech is 64,000 bits per second (b/s), but this rate is reduced to 2400 b/s for transmission over narrowband channels (having a bandwidth of approximately 3 KHz). The recently developed narrowband linear predictive coder (LPC) operating at 2400 b/s is such an example, and it is expected to be deployed extensively in the near future. The 2400-b/s LPC has been standardized within government agencies (Federal Standard 1015 or MIL-STD-188-113), and it has been adopted by NATO allied forces (STANAG 4198).

Recently, however, data rates above and below 2400 b/s have been gaining considerable interest; in particular, very-low-data-rate (VLDR) (i.e., 800 b/s or less) and 4800 b/s. Future research and development (R&D) efforts should be expanded in these areas according to a statement made by Mr. Donald C. Latham, Deputy Under Secretary of Defense (Communications, Command, Control, and Intelligence) which was related by the chairman of the Department of Defense (DoD) Digital Voice Processor Consortium on January 31, 1984.

There is a real need for voice processors operating at these data rates. The VLDR voice processor is for specialized military voice communication systems where a reliable connectivity is critically dependent on the reduced speech information rate. The implementation of an 800-b/s voice processor is not an easy task because the voice processor eliminates approximately 99% of the bit rate associated with the original speech. The intelligibility scores measured by the diagnostic rhyme test (DRT) have been in the low 80s from all previous experimental 800-b/s voice processors. Improvement in intelligibility is definitely desired.

On the other hand, an improved voice processor operating at 4800 b/s is also useful. It is well known that the 2400-b/s LPC does not reproduce indistinct or rapidly spoken speech. It is also somewhat biased against female voices (DRT differential of approximately 5.5 points). In fact, the 2400-b/s LPC is a difficult device to talk over, according to recent communicability tests conducted at the Naval Research Laboratory (NRL) [1]. As illustrated in Fig. 1, communicability score for the 2400-b/s LPC lags significantly behind that of a 9600-b/s LPC. The use of a 9600-b/s voice processor, however, may not be an ideal solution for all narrowband users because some narrowband channels cannot support a data rate of 9600 b/s. A preferred solution would be to use 4800 b/s, a data rate not far above 2400 b/s. It is interesting to recall that early experimental 2400-b/s LPCs developed in the mid-1970s routinely incorporated the 4800-b/s mode (and in some cases the 3600-b/s mode as well) to provide improved speech quality at the expense of a slightly higher data rate. For one reason or another, such an option was gradually dropped in the recently developed 2400-b/s LPC. This was an unfortunate oversight in retrospect.

During the development of these voice processing algorithms, we have made considerable effort to minimize computational complexity in order to make real-time operation feasible using present-day

Manuscript approved August 8, 1984.

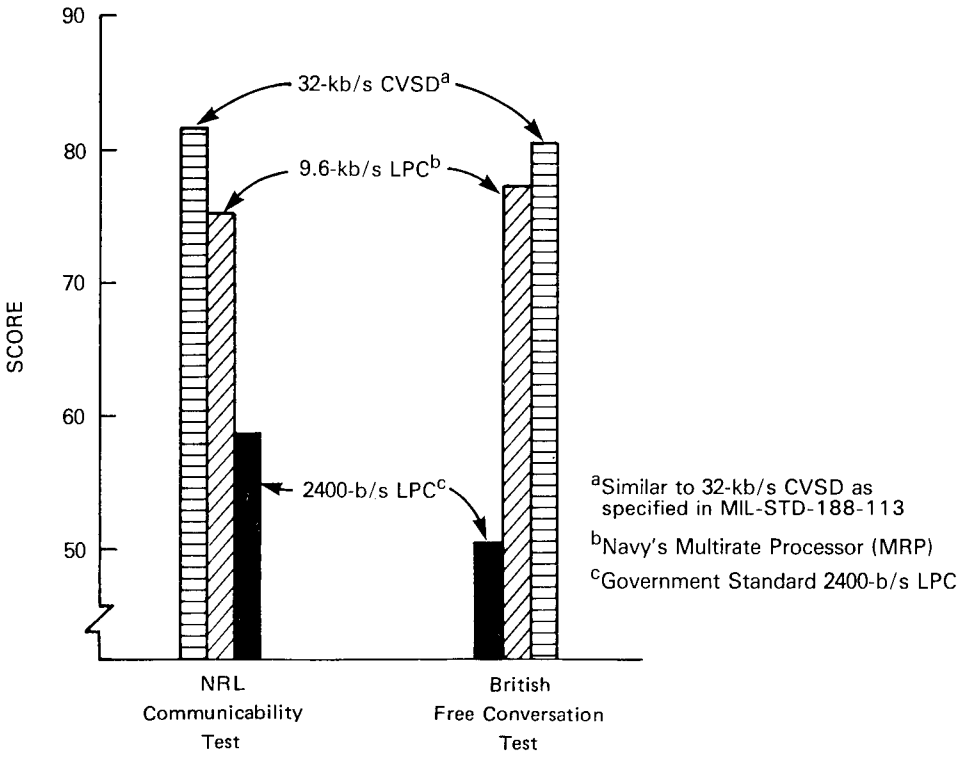


Fig. 1 — Two conversational test scores for various voice processors. The lower the score, the more effort is needed to communicate. This figure implies that the 2400-b/s LPC is not an easy device to talk over. The telephone score is in the lower 90s. In the NRL Communicability Test, the subjects' task is an abbreviated version of the pencil-and-paper game "battleship." In the British Free Conversation Test, subjects are given some task such as the comparison of pairs of photographs that induces the participants to talk for about ten minutes.

hardware. Listening tests alone cannot evaluate fully the actual usability of a voice processor in a two-way communication link. Only conversational tests (which require two real-time processors) allow the users to let each other know when communication has failed. The voice processing algorithms described in this report are well within real-time implementation using existing Navy-owned special signal processors, but real-time simulations have not yet been performed.

The Navy relies heavily on narrowband channels for voice communication. Because this capability is vital to the Navy, Naval Research Laboratory has been performing R&D on narrowband voice processors. In 1973, NRL developed one of the first narrowband LPCs capable of running in real time. Since 1975, NRL has investigated and sponsored several different speech encoding techniques designed to operate at 600 to 800 b/s. In 1981, NRL and Motorola produced a miniaturized 2400-b/s LPC that is only slightly larger than a standard desk telephone. During 1982 and 1983, NRL conducted studies to improve the 2400-b/s LPC without altering the interoperability requirements specified by Federal-Standard 1015. Now, a study has been made to implement both 800- and 4800-b/s voice processors. This report is a result of continuing efforts by NRL to make narrowband voice processors more acceptable to general users with diversified operating conditions.

BACKGROUND

Since Dudley invented what is known as the vocoder (derived from *voice coder* [2]) some 40 years ago, the fundamental principle employed by the narrowband voice encoder has not significantly changed. As illustrated in Fig. 2, the speech synthesizer consists of a filter representing the vocal tract

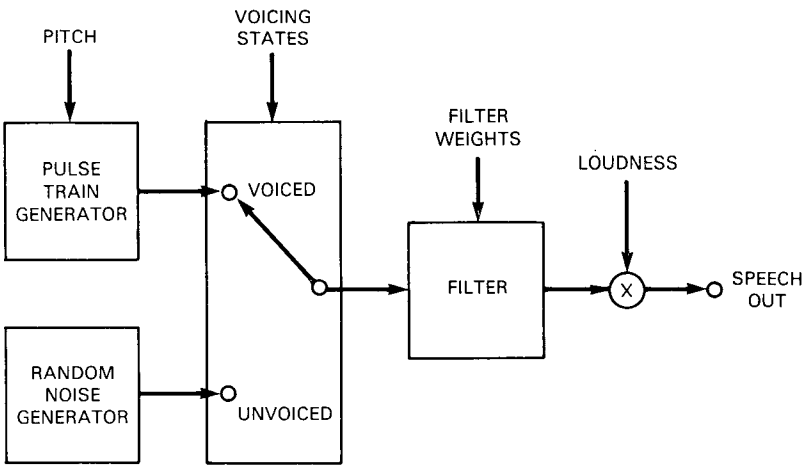


Fig. 2 — Simplified electrical analogue of speech generation by a narrowband voice processor. To generate continuous speech, the parameters (pitch period, voicing states, filter weights, and loudness) are updated 40 to 50 times per second.

and an excitation source that drives the filter. The excitation signal is usually one of two signals: random noise for the production of unvoiced sounds (i.e., consonants), or a pulse train for the production of voiced sounds (i.e., vowels). Such an excitation signal is still in use in current narrowband LPCs although there have been several different filters implemented.

The original device by Dudley, the channel vocoder, used ten contiguous narrowband-bandpass filters [2]. Subsequent channel vocoders most commonly used 16 channels, but some had as many as 19 channels. Well-designed channel vocoders achieved DRT scores in the upper 80s at 2400 b/s. The channel vocoder resistance to transmission-bit errors is remarkable because an error in a single filter parameter results in synthesized speech distortion only in that particular frequency band. According to previous tests, processed speech retains an acceptable speech intelligibility (i.e., DRT score of 81) even with 5% random-transmission-bit errors. Consonant sounds are rather realistic for certain channel vocoders because the passband extends up to 7 kHz or more. But, a major limitation of the channel vocoder in general is its unnatural vowel sounds somewhat akin to that of speech sounds propagated through a long hollow pipe.

Speech quality is greatly improved if the amplitude response of the filter shown in Fig. 2 models directly the speech-spectral envelope. Such a vocoder has been devised and is known as the spectral-envelope-estimation vocoder [3]. In this vocoder, speech samples are converted to the spectral envelope using a 256-point (FFT), and the resulting log-spectral envelope is down-sampled by a factor of 3. Thus, the speech-spectral envelope is represented by $128/3 \approx 42$ points.

A significant amount of data-rate reduction is achieved by the formant vocoder which transmits speech-spectral characteristics only near the resonant frequencies. The vocal-tract filter may be represented by three or more finite-Q resonators connected either in series or parallel. If the resonators are connected in series, relative formant amplitudes are not required [4,5]. This is similar to an all-pole filter that does not require information about the individual pole residue as is discussed below. If the resonators are connected in parallel, relative formant amplitudes are required to adjust channel gains [6]. In either configuration, formant bandwidth may be transmitted, or it may be assigned at the receiver causing degraded speech but achieving a lower bit rate.

Since the early 1970s, an all-pole filter has been used more extensively as the vocal-tract filter in narrowband voice processors [7]. The notable example is the current narrowband LPC. One advantage of using an all-pole filter is that its phase response is a continuous function of frequency (unlike that of

a continuous-filter bank employed by the channel vocoder). In addition, the amplitude response of an all-pole filter is rather faithful to the speech-spectral envelope. As an added advantage, both the filter coefficient generation and speech synthesis are computationally efficient. Unfortunately, a major drawback is a lack of robustness under a bit-error condition. An error in any one filter coefficient causes speech spectral distortions over the entire passband, not only in resonant frequencies but also in resonant amplitude. Note that resonant amplitudes of an all-pole filter are specified implicitly by resonant frequencies. A 5% random-bit error can reduce a DRT score by as much as 22 points, in contrast to only 7 points for the channel vocoder. Thus, some form of forward-error protection is desirable in LPC, as provided in the government standardized narrowband LPC.

We can represent an all-pole filter in many ways. The most direct representation is a positive feedback loop with a transversal filter in the feedback loop (see the appendix). In this representation, the filter coefficients are prediction coefficients. Prediction coefficients are the weighting factors appearing in the basic prediction equation; namely, a time sample is expressed as a weighted sum of past samples. These coefficients are actually not well suited for transmission because a bit error in any one coefficient can cause the synthesis filter to become unstable. Although there are many ways of checking the filter stability at the receiver (i.e., Hurwitz-Routh criterion and Schur-Cohn criterion are among the best known [8]), they all need a fair amount of computation. In retrospect, one of the most significant factors contributing to a successful implementation of the current narrowband LPC was the choice in the early 1970s to transmit reflection coefficients rather than prediction coefficients.

Reflection coefficients are transformed prediction coefficients which are also the coefficients of an all-pole filter represented by a cascaded-lattice filter (see the appendix). The advantage of transmitting reflection coefficients is that the stability of the synthesis filter is assured if the magnitude of each coefficient is between plus and minus one. In fact, the synthesis filter of the narrowband LPC never becomes unstable because the coefficient coding/decoding tables do not yield coefficients which can give rise to filter instability. The weakness of reflection coefficients is that a change in one coefficient causes speech spectral changes in the entire passband.

To overcome this weakness, we present another representation of an all-pole filter for use in both the 800- and 4800-b/s voice processors. In this all-pole filter representation, the parameters are line-spectrum frequencies (LSFs) (i.e., resonant frequencies with an infinite-Q or discrete frequencies). It is worthwhile to note that the vocal-tract-filter parameters are once again frequency-domain parameters as they are in the channel vocoder and formant vocoder. As depicted in Fig. 3, LSFs may be obtained from prediction coefficients via a transformation; similarly, reflection coefficients may be obtained from prediction coefficients via a transformation. An advantage of using LSFs is that the error in one LSF affects the synthesized spectrum near that frequency. Another advantage for using LSFs is that they may be more readily quantized in accordance with properties of auditory perception to save bits (i.e., coarser quantization of the higher frequency spectral components).

Prediction coefficients may be transformed into LSFs through the decomposition of the pulse response of the LPC-analysis filter into even- and odd-time sequences (Fig. 4). This decomposition is reversible because the original pulse response can be obtained by half the sum of the even- and odd-time sequences. As will be shown, the even-time sequence has roots along the unit circle in the complex plane. Thus the even-time sequences may be represented by LSFs. Likewise, the odd-time sequence has roots along the unit circle in the complex plane. Hence, the odd-time sequence may also be represented by LSFs.

Line-spectrum representation of prediction coefficients was first made public by F. Itakura in 1975 at the 89th meeting of the Acoustic Society of America [9]. Since then, applications of LSFs have been pursued mainly in Japan. It is significant to note that among 17 different speech synthesis chips which Japan produced during the past several years, one of them—ECL-1565—uses LSFs as filter parameters

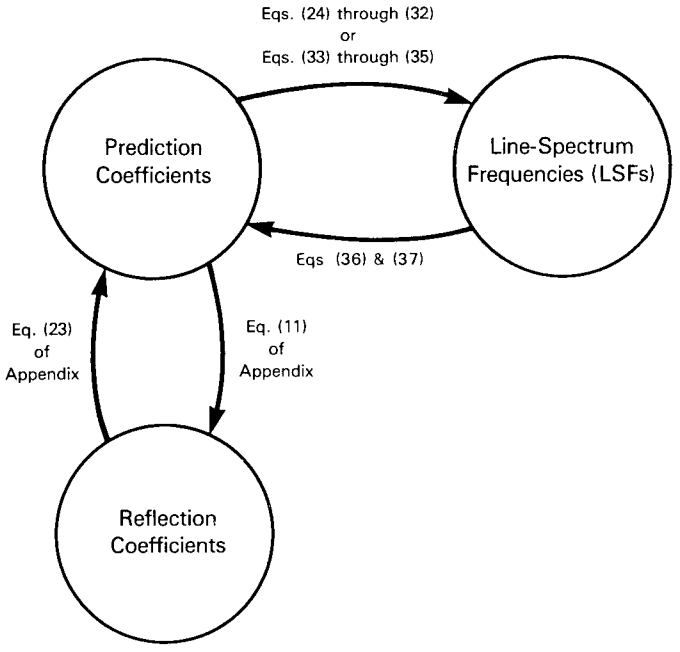
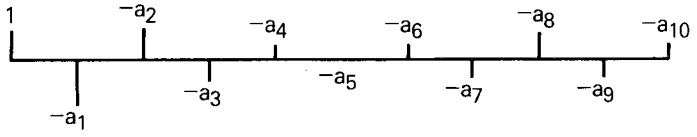
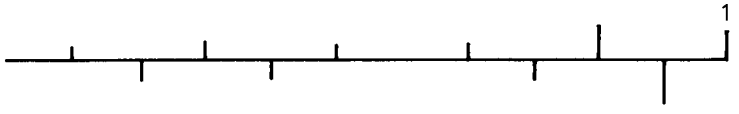


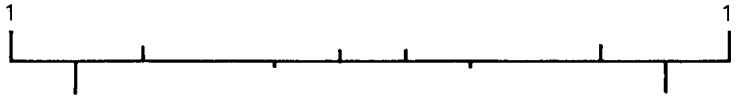
Fig. 3 — Parameters related to linear predictive analysis. When a time sample is expressed as a linear combination of past samples, the weighting factors are prediction coefficients (see the appendix). Prediction coefficients are never transmitted as speech parameters because bit errors can cause the synthesis filter to become unstable. Reflection coefficients are filter parameters which may be obtained directly from speech samples or from prediction coefficients (see the appendix). The synthesis filter is stable as long as the magnitude of each reflection is confined between plus and minus one. Currently, all LPC-based voice processors transmit reflection coefficients. LSFs and prediction coefficients are mutually transformable. Like reflection coefficients, LSFs have their own unique properties to consider for optimum quantization.



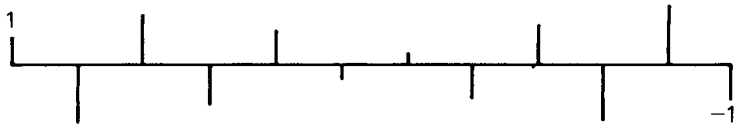
(a) Pulse response of tenth-order LPC analysis filter in which a_1 through a_{10} are prediction coefficients



(b) Time-shifted and time-reversed waveform of (a)



(c) Sum of waveforms: (a) plus (b). The time sequence is even symmetric with respect of its midpoint. All roots of this sequence are along the unit circle of the complex plane, with a real root at −1. Note that the first and last samples are both 1.



(d) Difference of waveforms: (a) minus (b). The time sequence is odd symmetric with respect to its midpoint. All roots of this sequence are also along the unit circle of the complex plane, with a real root at 1.

Fig. 4 — The response of the LPC-analysis filter decomposed into even- and odd-time sequences. This decomposition is the basis of the transformation from prediction coefficients to LSFs. The amplitude spectra of these waveforms are shown in Figs. 7 and 8.

[10]. In the United States, however, very little work has been done in this area. Only recently, a conference paper [11] and an article in a periodical [12] have been published related to this topic.

Our report also explores the application of LSFs. We emphasize the implementation of an 800-b/s pitch-excited LPC and a 4800-b/s nonpitch-excited LPC. Synthesized speech derived from LSFs using these two implementations was evaluated for the first time by formalized test procedures using the DRT. In addition, this report contains the derivation of all the necessary equations, results obtained from an investigation of the parameter sensitivities on spectral distortions, and outcome of a perceptual experiment using sounds generated from LSFs.

LSFs AS FILTER PARAMETERS

Currently, the most frequently used parameters for the LPC-analysis and synthesis filters are prediction coefficients or reflection coefficients. This section derives LSFs which are equivalent filter parameters.

LPC-Analysis Filter

LPC-analysis filter transforms speech samples into prediction residual samples. The most commonly used filter parameters have been either prediction coefficients or reflection coefficients (see the appendix). A functionally equivalent LPC-analysis filter may be constructed from the sum of two filters with even and odd symmetries. The basic principle is similar to the decomposition of an arbitrary function into a sum of even and odd functions [13].

We can express the transfer function of the nth-order LPC-analysis filter as

$$A_n(z) = 1 - \alpha_1 z^{-1} - \alpha_2 z^{-2} - \ldots - \alpha_n z^{-n}, \tag{1}$$

where α_i is the ith prediction coefficient of the nth-order predictor (i.e., α_i is a simplified notation of $\alpha_{i|n}$ used in the appendix), and z^{-1} is a one-sample delay operator. The recursive relationship of $A_{n+1}(z)$ in terms of $A_n(z)$, as noted in the appendix, is

$$A_{n+1}(z) = A_n(z) - k_{n+1} z^{-(n+1)} A_n(z^{-1}), \tag{2}$$

where k_{n+1} is the $(n + 1)$th reflection coefficient which equals α_{n+1} of the $(n + 1)$th-order predictor.

Let $P_{n+1}(z)$ be $A_{n+1}(z)$ with $k_{n+1} = 1$ (i.e., an open-end termination). Thus,

$$P_{n+1}(z) = A_n(z) - z^{-(n+1)} A_n(z^{-1}). \tag{3}$$

Likewise, let $Q_{n+1}(z)$ be $A_{n+1}(z)$ with $k_{n+1} = -1$ (i.e., a closed-end termination). Thus,

$$Q_{n+1}(z) = A_n(z) + z^{-(n+1)} A_n(z^{-1}). \tag{4}$$

From Eqs. (3) and (4), we express $A_n(z)$ as

$$A_n(z) = \frac{1}{2}[P_{n+1}(z) + Q_{n+1}(z)]. \tag{5}$$

Equation (5) is an alternative form of the LPC-analysis filter. The nature of $P_{n+1}(z)$ and $Q_{n+1}(z)$ is yet to be discussed.

$P_{n+1}(z)$ of Eq. (3) may be considered as the pulse response of a *difference filter* which takes the difference between the nth-order LPC-analysis filter output and its conjugate filter output. Note that $P_{n+1}(z)$ is odd-symmetric with respect to its midpoint (Fig. 4). On the other hand, $Q_{n+1}(z)$ of Eq. (4) may be considered as the pulse response of a *sum filter* which takes the sum of the $A_n(z)$ output and its conjugate filter output. Note that $Q_{n+1}(z)$ is even-symmetrtic with respect to its midpoint (Fig. 4).

This kind of filter composition need not be associated with the speech analysis or the inverse vocal-tract filter. The decomposition expressed by Eq. (5) holds for any arbitrary finite-duration-impulse-response filter. Similar decomposition has been exploited in the stability study of linear systems [8,14].

To show that all roots of $P_{n+1}(z)$ have roots along the unit circle in the z-plane, $A_n(z)$ expressed by Eq. (1) is substituted for $A_n(z)$ in Eq. (3). Thus,

$$P_{n+1}(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_{n/2}\, z^{-n/2} \underbrace{\qquad\qquad}_{\left(\frac{n}{2}+1\right)\text{ terms}} \underbrace{-a_{n/2}z^{-(n/2)+1} - \ldots - a_2 z^{-n+1} - a_1 z^{-n} - z^{-(n+1)}}_{\left(\frac{n}{2}+1\right)\text{ terms}}, \tag{6}$$

where

$$\begin{aligned} a_1 &= -\alpha_1 + \alpha_n \\ a_2 &= -\alpha_2 + \alpha_{n-1} \\ &\cdot \\ &\cdot \\ &\cdot \\ a_{n/2} &= -\alpha_{n/2} + \alpha_{(n/2)-1}\,. \end{aligned} \tag{7}$$

Since $P_{n+1}(z)$ has an odd symmetry with respect to its center coefficient, $z = 1$ is a root (this real root is an artifact introduced into the decomposition of $A_n(z)$). Thus we may factor Eq. (6) into

$$P_{n+1}(z) = \underbrace{(1 - z^{-1})[1 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_{(n/2)-1} z^{-(n/2)-1}}_{\left(\frac{n}{2}\text{ terms}\right)} \underbrace{+ d_{n/2} z^{-n/2} + \ldots + b_2 z^{-n+2} + b_1 z^{-n+1} + z^{-n}]}_{\left(\frac{n}{2}\text{ terms}\right)}. \tag{8}$$

The expression inside the brackets of Eq. (8) has a well-structured form: (a) it is even symmetric with respect to the center coefficient, and (b) both the first and last coefficients are unity. If the center coefficient was always positive and the largest value among the coefficients, the expression inside the brackets in Eq. (8) would be the result of a convolution of two identical filter responses. Since such is not the case, this expression is a product of vectors each having a unit amplitude. Thus, $P_{n+1}(z)$ of Eq. (8) may be expressed as

$$\begin{aligned} P_{n+1}(z) &= (1 - z^{-1}) \prod_{k=1}^{n/2} (1 - \epsilon^{j\theta_k} z^{-1})(1 - \epsilon^{-j\theta_k} z^{-1}) \\ &= (1 - z^{-1}) \prod_{k=1}^{n/2} (1 + d_k z^{-1} + z^{-2}), \end{aligned} \tag{9}$$

where

$$\begin{aligned} d_k &= -2 \cos(\theta_k), \qquad 0 \leqslant \theta_k \leqslant \pi, \\ &= -2 \cos(2\pi f_k t_s), \end{aligned} \tag{10}$$

in which f_k is the kth LSF in hertz associated with $P_{n+1}(z)$, and t_s is the speech sampling time interval.

Likewise, to show that all roots of $Q_{n+1}(z)$ have roots along the unit cricle in the z-plane, $A_n(z)$ expressed by Eq. (1) is substituted for $A_n(z)$ in Eq. (4). Thus,

$$\begin{aligned} Q_{n+1}(z) = 1 &+ c_1 z^{-1} + c_2 z^{-2} + \ldots + c_{n/2} z^{-n/2} \\ &+ c_{n/2} z^{-(n/2)+1} + \ldots + c_2 z^{-n+1} + c_1 z^{-n} + z^{-(n+1)}, \end{aligned} \tag{11}$$

where

$$\begin{aligned} c_1 &= -\alpha_1 - \alpha_n \\ c_2 &= -\alpha_2 - \alpha_{n-1} \\ &\cdot \\ &\cdot \\ &\cdot \\ c_{n/2} &= -\alpha_{n/2} - \alpha_{(n/2)-1}. \end{aligned} \tag{12}$$

Since $Q_{n+1}(z)$ has an even symmetry with respect to its center coefficient, $z = -1$ is a root (this real root is an artifact introduced in the decomposition of $A_n(z)$). Thus, we can factor Eq. (11) into

$$\begin{aligned} Q_{n+1}(z) = (1 + z^{-1})[1 &+ d_1 z^{-1} + d_2 z^{-2} + \ldots + d_{(n/2)-1} z^{-(n/2)-1} \\ &+ d_{n/2} z^{-n/2} + \ldots + d_2 z^{-n+2} + d_1 z^{-n+1} + z^{-n}], \end{aligned} \tag{13}$$

where

$$\begin{aligned} d_1 &= -1 + c_1 \\ d_2 &= 1 - c_1 + c_2 \\ d_3 &= -1 + c_1 - c_2 + c_3. \\ &\cdot \\ &\cdot \\ &\cdot \end{aligned} \tag{14}$$

The quantity inside the brackets of Eq. (13) has a similar form to that of $P_{n+1}(z)$ in Eq. (8). Thus, $Q_{n+1}(z)$ may be factored as

$$Q_{n+1}(z) = (1 + z^{-1}) \prod_{k=1}^{n/2} (1 + d'_k z^{-1} + z^{-2}), \tag{15}$$

where

$$d'_k = -2 \cos(2\pi f'_k t_s), \tag{16}$$

and f'_k is the kth LSF associated with $Q_{n+1}(z)$. Figure 5 is a block diagram of the LPC-analysis filter in which filter parameters are LSFs.

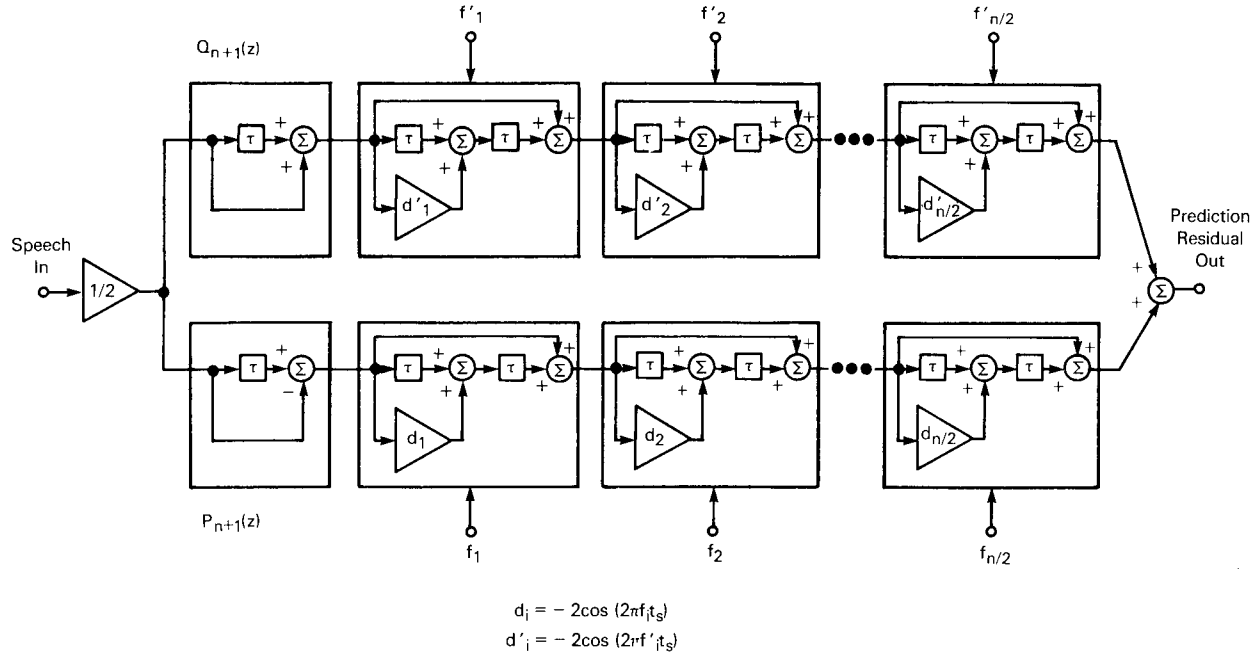


Fig. 5 — Block diagram of an nth-order LPC-analysis filter with LSFs as filter weights (n is even). This filter is functionally identical to the LPC-analysis filters with prediction coefficients or reflection coefficients as filter weights (see the appendix). As discussed later, LSFs are naturally ordered with $f_1 < f'_1 < f_2 < f'_2 \ldots$.

LPC-Synthesis Filter

The LPC-synthesis filter is the inverse of the LPC-analysis filter. Thus, the transfer function of the LPC-synthesis filter, $H_n(z)$, is

$$\begin{aligned} H_n(z) &= \frac{1}{A_n(z)} \\ &= \frac{1}{\frac{1}{2}[P_{n+1}(z) + Q_{n+1}(z)]} \\ &= \frac{1}{1 - \frac{1}{2}[1 - P_{n+1}(z)] + \frac{1}{2}[1 - Q_{n+1}(z)]}. \end{aligned} \tag{17}$$

Equation (17) is the form of a positive feedback in which the feedback element is the quantity inside the bracket. Expressing this feedback element in terms of LSFs, we obtain

$$\begin{aligned} F(z) &= \frac{1}{2}\left\{[1 - P_{n+1}(z)] + [1 - Q_{n+1}(z)]\right\} \\ &= \frac{1}{2}\left\{\left[1 - (1 - z^{-1})\prod_{k=1}^{n/2}(1 + d_k z^{-1} + z^{-2})\right] + \left[1 - (1 + z^{-1})\prod_{k=1}^{n/2}(1 + d'_k z^{-1} + z^{-2})\right]\right\} \\ &= \frac{1}{2}\left\{\left[1 - \prod_{k=1}^{n/2}(1 + d_k z^{-1} + z^{-2})\right] + z^{-1}\prod_{k=1}^{n/2}(1 + d_k z^{-1} + z^{-2})\right. \\ &\quad \left. + \left[1 - \prod_{k=1}^{n/2}(1 + d_k z^{-1} + z^{-2})\right] - z^{-1}\prod_{k=1}^{n/2}(1 + d'_k z^{-1} + z^{-2})\right\}. \end{aligned} \tag{18}$$

In Eq. (18), both the second and fourth terms may be realized by cascading second-order filters identical to those appearing in the LPC-analysis filter shown by Fig. 5. The first and third terms (which do not appear in the LPC-analysis filter) may be implemented more economically if the following recursive relationship is exploited. Let $G_k(z)$ represent either the first or the third term of Eq. (18). Thus,

$$G_K(z) = 1 - \prod_{k=1}^{K}(1 + g_k z^{-1} + z^{-2}), \qquad K \leqslant n/2. \tag{19}$$

$G_K(z)$ in terms of $G_{K-1}(z)$ may be expressed as

$$G_K(z) = -z^{-1}(g_K + z^{-1}) \prod_{k=1}^{K-1}(1 + g_K z^{-1} + z^{-2}) + G_{K-1}(z). \tag{20}$$

Note that the first term of Eq. (20) without z^{-1} is available from each section of the second-order filter (output of the first summer in each of the heavy-lined boxes in Fig. 6). Thus, the total operations required to compute either the first or third terms of Eq. (18) are $n/2$ summations. The LPC-synthesis filter is shown in Fig. 6.

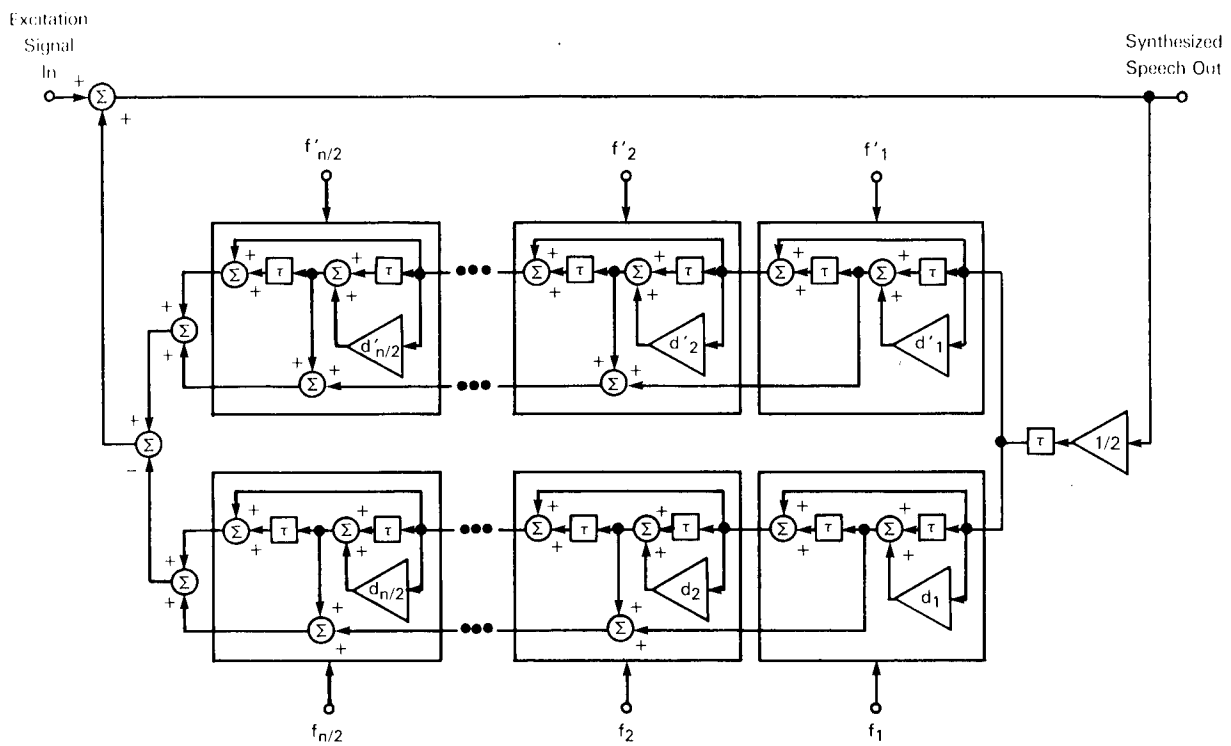


Fig. 6 — An nth-order LPC-synthesis filter with LSFs as filter weights (n is even). This filter is functionally identical to the LPC-synthesis filters with prediction coefficients or reflection coefficients as filter weights (see the appendix).

FILTER PARAMETER TRANSFORMATIONS

As indicated in Fig. 3, we can transform a set of prediction coefficients into a set of LSFs, and vice versa. This section presents these conversion algorithms.

Conversion of Prediction Coefficients to LSFs

Conversion of prediction coefficients to LSFs is, in essence, finding roots of the difference and sum filters expressed previously as Eqs. (3) and (4):

$$P_{n+1}(z) = A_n(z) - z^{-(n+1)}A_n(z^{-1}) \quad \text{(difference filter)}, \tag{3}$$

$$Q_{n+1}(z) = A_n(z) + z^{-(n+1)}A_n(z^{-1}) \quad \text{(sum filter)}. \tag{4}$$

Since the roots of either $P_{n+1}(z)$ or $Q_{n+1}(z)$ lie only along the unit circle of the z-plane, the use of generalized root-finding procedures is more than what is required for the present application. A preferred approach is a numerical method for determining the frequencies which make the amplitude response of either the difference or sum filter zero. In this approach, both $P_{n+1}(z)$ and $Q_{n+1}(z)$ are evaluated for various values for z (where z is $\exp(j2\pi ft_s)$ and $j = \sqrt{-1}$ from a frequency of 0 Hz to the upper cutoff frequency. This form of computation is systematic; hence, the necessary software is relatively compact. Furthermore, the maximum amount of computation needed to find all the LSFs is fixed, contrary to those methods which rely on the convergence of the solution (viz., Newton's method). Information regarding the maximum number of computational steps is important for the implementation of the algorithm in a real-time voice processor.

If the difference and sum of two quantities contain all the necessary information, then the ratio of the same two quantities also contains the same information. The use of the ratio filter, an alternative approach for finding LSFs, is based on the rearranged expression for $P_{n+1}(z)$ and $Q_{n+1}(z)$:

$$P_{n+1}(z) = A_n(z)[1 - R_{n+1}(z)], \tag{21}$$

$$Q_{n+1}(z) = A_n(z)[1 + R_{n+1}(z)], \tag{22}$$

where

$$R_{n+1}(z) = \frac{z^{-(n+1)}A_n(z^{-1})}{A_n(z)} \quad \text{(ratio filter)}. \tag{23}$$

The ratio filter is an all-pass filter (i.e., a phase shifter with a flat amplitude response). When the phase angle of the ratio filter is a multiple of π radians, the amplitude response of the difference filter is zero. On the other hand, when the phase angle of the ratio filter is zero (or a multiple of 2 radians), the amplitude response of the sum filter is zero. Hence, frequencies which give rise to these two phase angles are the LSFs. We describe in detail the computational steps for these two approaches.

Approach 1: Using Amplitude Response of Sum and Difference Filters

We begin computation with a set of prediction coefficients generated by LPC analysis using one frame of speech samples (approximately 100 to 200 samples).

(a) The coefficients of the LPC-analysis filter transfer function are generated from

$$a_n(1) = 1, \tag{24a}$$

$$a_n(i) = -\alpha_{i-1}, \; i = 2, 3, \ldots, n + 1, \tag{24b}$$

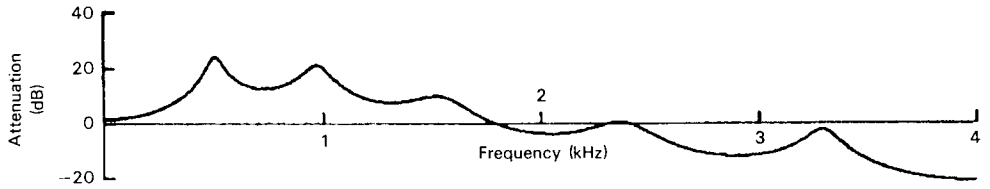
where $a_n(i)$ is the ith coefficient of $A_n(z)$ and i is the ith prediction coefficient. The frequency response of $A_n(z)$ is illustrated in Figs. 7(a) and 8(a).

(b) The coefficients of the difference filter are generated by

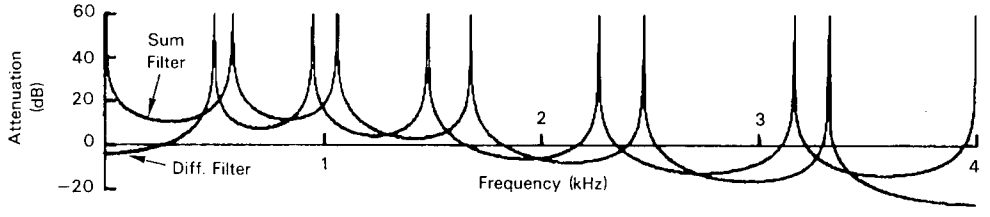
$$P_{n+1}(1) = 1, \tag{25a}$$

$$p_{n+1}(i) = a_n(i) - a_n(n + 3 - i), \; i = 2,3,\ldots, n + 2, \tag{25b}$$

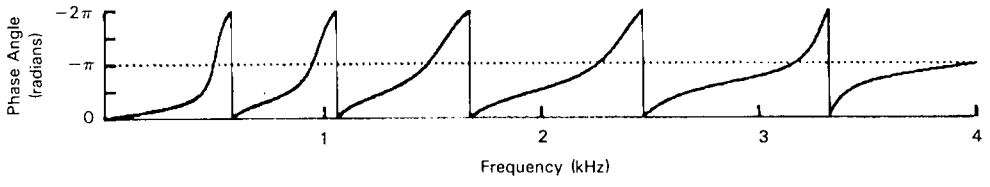
where $p_{n+1}(i)$ is the ith coefficient of $P_{n+1}(z)$. The frequency response of $P_{n+1}(z)$ is shown in Figs. 7(b) and 8(b).



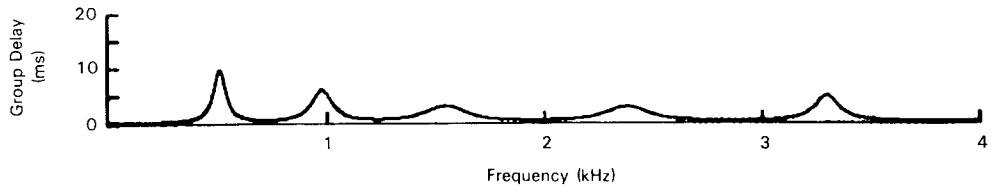
(a) Amplitude response of LPC-analysis filter



(b) Amplitude responses of sum and difference filters

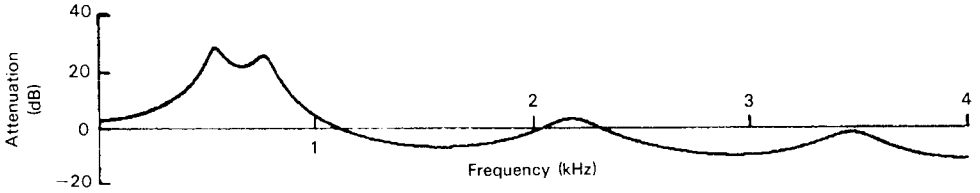


(c) Phase response of ratio filter

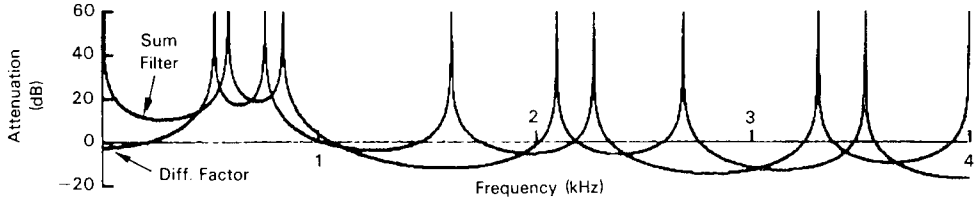


(d) Group delay of ratio filter

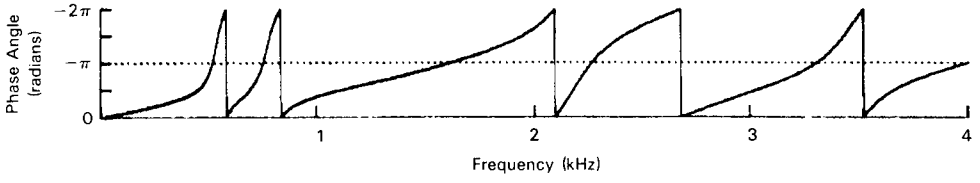
Fig. 7 — Frequency response of the LPC-analysis filter and other derivative filters containing LSF information (Example #1). Figure 7(a) is the tenth-order LPC-analysis-filter-amplitude response associated with speech sound /o/ in "strong." The LPC-synthesis-filter-amplitude response is identical to Fig. 7(a) if the vertical axis is labeled as "Gain (dB)" instead of "Attenuation (dB)." LPC-synthesis-filter-amplitude response is an all-pole approximation to the speech-spectral envelope. Figure 7(b) is the amplitude response of the difference and sum filters as defined in the text. The null frequencies associated with either filter are LSFs. To see the effect, a real pole has not been removed in either the difference or sum filter. The expression "line spectrum" originated from the shapes of these frequency responses. Figure 7(c) is the phase response of the ratio filter defined in the text. Note that when the phase response becomes π radians, the amplitude response of the difference filter becomes null. On the other hand, when the phase response becomes 0 or 2 radians, the amplitude response of the sum filter becomes null. Figure 7(d) is the group delay of the ratio filter. The group delay is large near speech-resonant frequencies, and LSFs are close together.



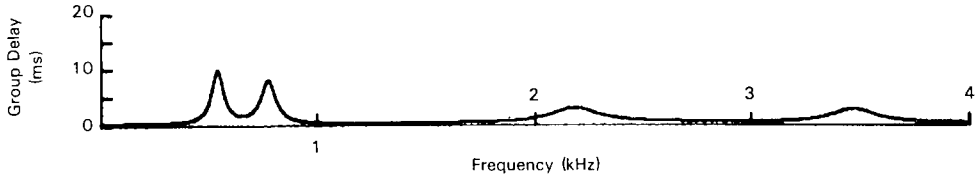
(a) Amplitude response of LPC-analysis filter



(b) Amplitude responses of sum and difference filters



(c) Phase response of ratio filter



(d) Group delay of ratio filter

Fig. 8 — Frequency responses of the LPC-analysis filter and other derivative filters containing LSF information (Example #2). Figure 8(a) is the tenth-order LPC-analysis-filter-amplitude response derived from speech /o/ in "oak." Figure 8(b) is the amplitude response of the difference and sum filters (i.e., the line spectrum). It is significant to note that some of the LSFs are not located near speech-resonant peaks because the number of LSF pairs (five pairs excluding extraneous ones) is greater than the number of speech-resonant peaks (four). In example #1, these two numbers are equal, and all the LSFs are located near speech-resonant peaks. As will be discussed, isolated LSFs are less sensitive to the amplitude spectrum. Figures 8(c) and 8(d) are the phase response and group delay of the ratio filter. Again, group delay is much larger near speech-resonant peaks.

(c) Likewise, coefficients of the sum filter are generated by

$$q_{n+1}(1) = 1, \tag{26a}$$

$$q_{n+1}(i) = a_n(i) + a_n(n + 3 - i), \quad i = 2,3,\ldots, n + 2, \tag{26b}$$

where $q_{n+1}(i)$ is the ith coefficient of $Q_{n+1}(z)$. The frequency response of $Q_{n+1}(z)$ is shown in Figs. 7(b) and 8(b).

(d) Since a real pole at $z = 1$ is an artifact introduced in the generation of the difference filter, $(1 - z^{-1})$ may be factored out from $P_{n+1}(z)$ to simplify computations. Thus,

$$P_{n+1}(z) = (1 - z^{-1})P_n(z), \tag{27}$$

and the coefficient sequence of $p_n(z)$ is

$$p_n(1) = p_{n+1}(1), \tag{28a}$$

and

$$p_n(i) = p_{n+1}(i) + p_n(i - 1), \; i = 2,3,\ldots, n + 1. \tag{28b}$$

(e) Likewise, $(1 + z^{-1})$ may be factored out from the sum filter. Thus,

$$Q_{n+1}(z) = (1 + z^{-1})\, Q_n(z), \tag{29}$$

and the coefficient sequence of $Q_n(z)$ is

$$q_n(1) = 1, \tag{30a}$$

and

$$q_n(i) = q_{n+1}(i) - q_n(i - 1), \; i = 2,3,\ldots, n + 1. \tag{30b}$$

(f) Use of the autocorrelation sequence simplifies the amplitude spectral analysis because it needs only the cosine transform. The autocorrelation sequence of $p_n(i)$ is obtained from

$$\phi_{pp}(i) = \sum_{j=1}^{n+3-i} p_n(i)p_n(i + j - 1), \quad i = 1,2\ldots, n + 2. \tag{31}$$

(g) The power spectrum of $p_n(i)$ is obtained from

$$\Phi_{pp}(kf_s) = \frac{1}{2\pi}\left[\phi_{pp}(1) + 2\sum_{i=1}^{n+1} \phi_{pp}(i + 1)\cos(2\pi ikf_s t_s)\right] \quad k = 1,2,\ldots, \tag{32}$$

where t_s is the speech-sampling time interval, and f_s is the frequency step in hertz. The choice of frequency step size is a major concern because a finer step size increases computation, whereas a coarser step size introduces irreversible spectral distortions. According to intelligibility testing using DRT, frequency steps of 10 Hz did not degrade the intelligibility of synthesized speech. Note that the above computations are terminated upon finding $n/2$ LSFs.

(h) LSFs are the frequencies which make $\Phi_{pp}(w)$ local minima (see Figs. 7(b) or 8(b) for the amplitude response of the difference filter). For the nth-order LPC-analysis filter, there will be $n/2$ LSFs (and one more at the upper-cutoff frequency if step (d) is omitted).

(i) For the sum filter, steps (f) through (h) are repeated using $q_n(i)$ in place of $p_n(i)$. For the nth-order LPC-analysis filter, there will be again $n/2$ LSFs (and one more at 0 Hz if step (e) is omitted). Note that LSFs associated with the sum filter are always above the corresponding LSFs associated with the difference filter.

Approach 2: Using Phase Response of Ratio Filter

The most significant difference between this approach and the previous one is that the spectral analysis is performed only once, contrary to twice in the previous approach. This approach, however, requires both the cosine and sine transforms, whereas the previous approach needs only the cosine transform. Furthermore, this approach requires inverse-tangent operations. Thus, there is no advantage from a computational point of view. This approach, however, allows interpolation of LSFs for finer resolutions which was impossible in the previous one. Most important, this method provides more readily the group delay of the ratio filter. As will be shown, the group delay at an LSF is directly related to the spectral sensitivity of that particular LSF.

As noted from Eq. (23), the phase spectrum of the ratio filter needs the phase spectrum of the LPC-analysis filter, $A_n(z)$. The complex spectrum of the LPC-analysis filter is

$$A_n(kf_s) = \sum_{i=0}^{n} a_n(i+1) \exp(-j2\pi ikf_s t_s), \quad j = \sqrt{-1}, \tag{33}$$

where $a_n(i)$ is the ith coefficient of the nth-order LPC-analysis filters. Since $a_n(i) = 1$ and $a_n(i) = -\alpha_i$ where α_i is the ith prediction coefficient, we can write Eq. (24) as

$$A_n(kf_s) = \left[1 - \sum_{i=1}^{n} \alpha_i \cos(2\pi ikf_s t_s)\right] + j\left[\sum_{i=1}^{n} \alpha_i \sin(2\pi kf_s t_s)\right]. \tag{34}$$

From Eqs. (23) and (25), the phase spectrum of the ratio filter, denoted by $\phi(kf_s)$, is

$$\phi(kf_s) = -(n+1)(2\pi kf_s t_s) - 2\tan^{-1}\left\{\frac{\sum_{i=1}^{n} \alpha_i \sin(2\pi ikf_s t_s)}{1 - \sum_{i=1}^{n} \alpha_i \cos(2\pi ikf_s t_s)}\right\}, \quad k = 1,2,\ldots. \tag{35}$$

Figures 7(c) and 8(c) illustrate the phase spectrum of the ratio filter computed by Eq. (35). As stated previously, LSFs are the frequencies which give rise to a phase value of a multiple of either $-\pi$ or -2π radians. Since the phase spectrum is a monotonically decreasing function as frequency increases, LSFs can be interpolated for a finer resolution. In Eq. (35), the frequency step is 10 Hz as in the previous approach. Above computation is terminated upon finding n number of LSFs.

Given a set of prediction coefficients, the use of Eq. (35) is the most direct way for computing the phase spectrum of the ratio filter. As shown later in Eq. (45) phase spectrum of the ratio filter can also be computed (although not as conveniently) from a set of roots of the LPC-analysis filter.

Conversion of LSFs to Prediction Coefficients

We can convert LSFs to prediction coefficients by solving for the coefficients of the polynomial which represent the transfer function of the LPC-analysis filter, $A_n(z)$, in terms of LSFs. To begin with, $A_n(z)$ can be expressed as a sum of two factored terms by substituting Eqs. (9) and (15) into Eq. (5). Thus,

$$A_n(z) = \frac{1}{2}\left[(1 - z^{-1})\prod_{k=1}^{n/2}(1 + d_k z^{-1} + z^{-2}) + (1 + z^{-1})\prod_{k=1}^{n/2}(1 + d'_k z^{-1} + z^{-2})\right] \tag{36}$$

When the product terms are multiplied out, the resulting polynomial is in the form

$$A_n(z) = 1 + \beta_1 z^{-1} + \beta_2 z^{-2} + \ldots + \beta_n z^{-n}. \tag{37}$$

Comparing Eq. (37) term by term with Eq. (1) indicates that the ith prediction coefficient is $-\beta_i$ for $i = 1,2,\ldots, n$.

The coefficients of Eq. (37), hence the solution for the prediction coefficients, can be obtained numerically by computing the $(n + 1)$ samples from the LPC-analysis filter under the excitation of a single pulse (i.e., 1 followed by n number of zeros). The number of computational steps needed for this solution, as determined from either Eq. (36) or the block diagram of the LPC-analysis filter shown in Fig. 5, is $n(n + 1)$ multiplications and $(2n + 3)(n + 1)$ summations as listed in Table 1.

Table 1 — Total Number of Arithmetic Operations Required for Speech Synthesis and Parameter Conversion

	To Synthesize Speech Samples (Arithmetic Operations per Sample)			To Convert LSFs to Prediction Coefficients (Arithmetic Operations per Frame[a])
	Using Prediction Coefficients (Fig. A1)	Using LSFs (Fig. 6)	Differential	
Multiplications	n	n	0	$n(n + 1)$
Summations	n	$3n + 2$	$2n + 2$	$(2n + 3)(n + 1)$

[a] Frame is usually 160 to 200 samples.

This kind of parameter conversion reduces the number of computational steps in the speech-synthesis algorithm. In fact, similar parameter conversions are currently being performed in some narrowband LPCs in which received reflection coefficients are converted to prediction coefficients prior to speech synthesis. Table 1 lists the total number of arithmetic operations necessary to synthesize one speech sample by using two different filter coefficients: one using prediction coefficients and another using LSFs. The difference is $(2n + 2)$ summations per speech sample in favor of using prediction coefficients.

As a numerical example, let n be 10 (i.e., LPC-synthesis filter with 10 filter weights), and frame size be 180 samples. If the prediction coefficients are used in lieu of LSFs, 3960 summing operations are saved for each frame. On the other hand, parameter conversion requires 110 multiplications and 253 summations, but this conversion is needed only once per frame. Thus, converting LSFs to prediction coefficients is beneficial for reducing computational steps during speech synthesis.

PROPERTIES OF LSFs

This section discusses several properties associated with LSFs. Spectral sensitivity to LSFs is discussed in a separate section because of the importance this has for influencing filter-parameter quantization.

Naturally Ordered Frequency Indices

For the nth-order LPC-analysis filter, there are always $(n + 2)$ LSFs (two are extraneous frequencies which do not contain speech information (Table 2)). Furthermore, the $(n + 2)$ LSFs are naturally ordered such that the ith LSF in one instant of time remains the ith LSF for another instant of time. Therefore, trajectories of LSFs are continuous, and they do not intersect each other as shown in Fig. 9.

Table 2 — LSFs Derived from a Tenth-Order LPC-Analysis Filter. The first LSF, time invariant at 0 Hz, is extraneous. Likewise, the last LSF, time invariant at a half-sampling frequency, is also extraneous. Both are created by the filter-decomposition process defined by Eq. (5).

Cumulative LSF Indices (F_k)	LSF Indices of Difference Filter (f_k)	LSF Indices of Sum Filter (f'_k)	Phase Angle of Ratio Filter in Eq. (35)
0	0	0	0 Radians
1	1		$-\pi$
2		1	-2π
3	2		-3π
4		2	-4π
5	3		-5π
6		3	-6π
7	4		-7π
8		4	-8π
9	5		-9π
10		5	-10π
11	6	0	-11π

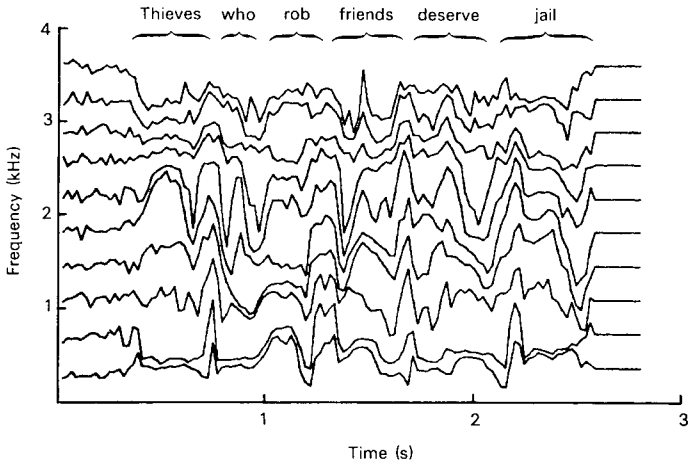


Fig. 9 — LSF trajectories computed from tenth-order LPC coefficients. LSFs do not disappear as do the formant frequencies, nor do they crisscross each other. LSFs are closer together near speech-resonant frequencies. On the other hand, LSFs become equally spaced if the input spectrum is flat (i.e., all prediction coefficients are zeros) as noted near the end of the LSF trajectories. For high-frequency-dominant-unvoiced sounds, all LSFs achieve higher values. These properties of LSFs are discussed in the next section.

To prove that the LSFs are naturally ordered is equivalent to proving that the phase angle of the filter is a monotonic function of frequency. As discussed previously in connection with the filter-parameter transformation from predicted coefficients to LSFs, LSFs are those frequencies which give rise to phase angles of the ratio filter of $0, -\pi, -2\pi, =3\pi$, etc. If the phase response of a filter is a monotonic function of frequency, then its first derivative with respect to frequency does not change its sign over the entire frequency domain.

To prove the above statement, the following three quantities are obtained from the transfer function of the ratio filter: frequency response, phase response, and group delay (i.e., first derivative of phase response). From Eq. (23), the transfer function of the ratio filter is

$$R_{n+1}(z) = \frac{z^{-(n+1)} A_n(z^{-1})}{A_n(z)} \quad \text{(ratio filter)}, \tag{23}$$

where $A_n(z)$ is the nth-order LPC-analysis-filter-transfer function. Equation (23) may be factored as

$$A_n(z) = \prod_{k=1}^{n} (1 - z_k z^{-1}), \tag{38}$$

where z_k is kth root of $A_n(z)$, and z_k lies inside the unit circle of the z-plane. Substituting Eq. (38) into Eq. (23) yields

$$\begin{aligned} R_{n+1}(z) &= z^{-(n+1)} \prod_{k=1}^{n} \frac{(1 - z_k{}^* z)}{(1 - z_k z^{-1})} \\ &= z^{-1} \prod_{k=1}^{n} \frac{(z^{-1} - z_k{}^*)}{(1 - z_k z^{-1})}, \end{aligned} \tag{39}$$

where z^* is a complex conjugate of z. Equation (39) may be represented as

$$R_{n+1}(z) = z^{-1} \prod_{k=1}^{n} \rho_k(z), \tag{40}$$

in which

$$\rho_k(z) = \frac{z^{-1} - z_k{}^*}{1 - z_k z^{-1}}. \tag{41}$$

In terms of a modulus and argument, the kth root of $A_n(z)$ is in the form of

$$z_k = r_k \epsilon^{-j\omega_k}, \quad j = \sqrt{-1}. \tag{42}$$

By substituting z for $\exp(j\omega t_s)$ and Eq. (42) into Eq. (41), the complex-frequency response of the ratio filter is achieved. Thus,

$$\rho_k(\omega) = \frac{1 - r_k \epsilon^{j(\omega - \omega_k) t_s}}{1 - r_k \epsilon^{-j(\omega - \omega_k) t_s}}. \tag{43}$$

A complex spectrum can be written in the form

$$\rho_k(\omega) = A_k(\omega) \epsilon^{j\phi_k(\omega)}. \tag{44}$$

The phase spectrum, $\phi_k(\omega)$, is the imaginary part of the logarithm of the complex spectrum. Thus,

$$\phi_k(\omega) = \text{Im}\,[\ln(\rho_k(\omega))]. \tag{45}$$

The group delay, denoted by $D_k(\omega)$, is by definition

$$D_k(\omega) = -\frac{d}{d\omega}\rho_k(\omega)$$
$$= -\mathrm{Im}\left[\frac{d}{d\omega}\ln(\phi_k(\omega))\right]. \tag{46}$$

Substituting Eq. (43) into Eq. (46) yields

$$D_k(\omega) = t_s\frac{1-r_k^2}{1-2r_k\cos(\omega-\omega_k)t_s+r_k^2}. \tag{47}$$

From Eq. (40), the total group delay is a sum of all partial delays and an additional one-sample delay. Thus, total group delay, denoted by $D(\omega)$, is

$$D(\omega) = t_s\left[1+\sum_{k=1}^{n}\frac{1-r_k^2}{1-2\,r_k\cos(\omega-\omega_k)t_s+r_k^2}\right]. \tag{48}$$

If all roots of the LPC-analysis filter lie inside the unit circle of the z-plane (i.e., $r_k < 1$ for all k's), the group delay expressed by Eq. (48) is positive for all frequencies regardless of the value of ω_k (i.e., arguments of the roots of $A_n(z)$, or speech-resonant frequencies). Hence, the phase response of the ratio filter is a monotonic function of frequency. As a result, the phase angles of $0,-\pi,-2\pi,-3\pi$, etc. appear in sequence as frequency increases. Thus, all LSFs are naturally ordered.

It is worthwhile to mention that the ratio filter expressed by Eq. (23) or (39) has application beyond computing LSFs. It has been used as a mapping function to transform one filter to another (viz., low-pass filter to high-pass filter, or low-pass filter to bandpass filter) [14,15]. In this application, the new variable may be defined as

$$Z^{-1} = z^{-1}\prod_{k=1}^{n}\frac{z^{-1}-z_k^*}{1-z_kz^{-1}}, \tag{49a}$$

or

$$Z^{-1} = \prod_{k=1}^{n}\frac{z^{-1}-z_k^*}{1-z_kz^{-1}}. \tag{49b}$$

This transformation maps the unit circle unto itself because

$$|Z|\begin{cases} < 1 & \text{for } |z| < 1 \\ = 1 & \text{for } |z| = 1. \\ > 1 & \text{for } |z| > 1 \end{cases} \tag{50}$$

In other words, the unit circle is invariant under this transformation.

Evenly Spaced Frequencies with Flat-Input Spectrum

If the input signal has a flat-amplitude spectrum for the entire passband, the resulting LSFs are evenly spaced, as illustrated in the far-right end of Fig. 9. For such an input signal, the transfer function of the LPC-analysis filter is unity (which means that the prediction residual is identical to the input). It follows from Eq. (38) that the moduli of all roots are zero (i.e., $r_k = 0$ for all k's). Hence, the group delay as obtained from Eq. (48) for this case is

$$D(\omega) = (n+1)t_s, \tag{51}$$

and the corresponding phase response is

$$\phi(\omega) = (n + 1)\,\omega t_s + \text{constant}. \tag{52}$$

Since the phase response is a linear function of frequency, phase angles of $0, -\pi, -2\pi, -3\pi \ldots$ occur at a fixed-frequency interval. Since LSFs are those frequencies associated with these phase angles, LSFs are evenly spaced. As noted from Eq. (52), the phase angle at 0 Hz is 0 radian, and the phase angle at the upper cutoff frequency is $(n + 1)\pi$ radians. Since LSFs occur at a multiple of π radians, the frequency separation between two adjacent LSFs is $\Omega/(n + 1)$ where Ω is the upper cutoff frequency.

This result may be obtained alternatively from the transfer functions of the difference and sum filters, $P_{n+1}(z)$ and $Q_{n+1}(z)$. For an input signal having a flat-amplitude spectrum over the entire passband, its autocorrelation coefficients are zero except for a delay of zero (i.e., an impulse function in the delay domain). Thus, prediction coefficients are zeros. Hence, as stated previously, the transfer function of the LPC-analysis filter is unity (i.e., $A_n(z) = 1$). Thus, transfer functions of the difference and sum filters, as obtained from Eqs. (3) and (4), are

$$P_{n+1}(z) = 1 - z^{-n+1}, \tag{53a}$$

and

$$Q_{n+1}(z) = 1 + z^{-(n+1)}. \tag{53b}$$

The solutions for these polynomials are well known because they are often quoted problems in complex-variable courses. Roots of these polynomials are:

$$\text{for } P_{n+1}(z)\text{: } z_k = \exp\left\{\frac{j2\pi}{(n+1)}k\right\}, \quad k = 1, 3, 5, \ldots, n + 1, \tag{54a}$$

$$\text{for } Q_{n+1}(z)\text{: } z_k = \exp\left\{\frac{j2\pi}{(n+1)}k\right\}, \quad k = 0, 2, 4, \ldots, n, \tag{54b}$$

where $j = \sqrt{-1}$. Thus, roots of the transfer functions of the difference and sum filters are interlaced, and the angular distance between any two interlaced roots (i.e., LSFs) is equal for the flat-input spectrum.

Closely Spaced Frequencies Near Input-Resonant Frequencies

As noted in Figs. 7 and 8, LSFs are closer together near input-resonant frequencies. This is because the group delay of the ratio filter near the input-resonant frequencies is larger than elsewhere.

The input-resonant frequencies are reflected in the roots of the LPC-analysis-filter-transfer function, $A_n(z)$. Let one root of $A_n(z)$ be $r_m \exp(j\omega_m)$. Group delay of the ratio filter at the input-resonant frequency, as obtained from Eq. (47), is

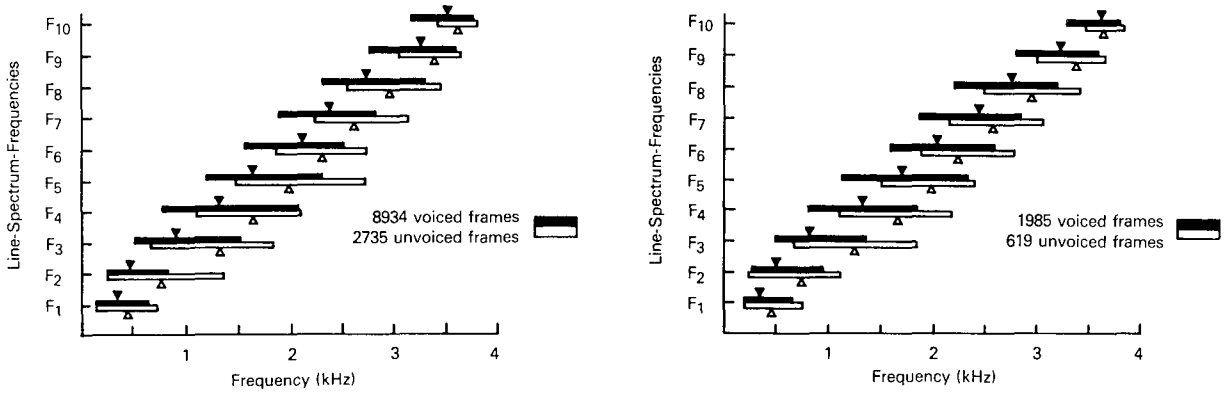
$$D(\omega_m) = t_s\left[1 + \left(\frac{1 + r_m}{1 - r_m}\right) + \sum_{\substack{k=1 \\ k \neq m}}^{n} \frac{1 - r_k^2}{1 - 2r_k \cos(\omega_m - \omega_k)\, t_s + r_k^2}\right], \quad 1 \leqslant m \leqslant n. \tag{55}$$

When r_m is near unity (as would be the case with resonant frequencies associated with vowels), the second term of Eq. (55) contributes mostly to the total delay. Thus, the group delay becomes relatively larger near speech-resonant frequencies, as well evidenced in Figs. 7(d) and 8(d). An increased group delay means a reduced-frequency interval during which the angle is decreased by π radians. Hence, the two adjacent LSFs are closer together.

Frequency Distributions

Each LSF has an excursion range which is dependent on the speech, speaker, and other factors such as the nature of the preemphasis prior to LPC analysis and low-frequency-cutoff characteristics of the front-end-audio circuit. Magnitude of the frequency range of each individual LSF is essential information for encoding them.

Figure 10 is a plot of LSF distributions computed from 54 male and 12 female speakers, each uttering two sentences. For each category of speaker, a separate frequency distribution plot was made for voiced and unvoiced frames (excluding nonvoiced or silent frames). Preemphasis is performed by a single-zero filter having a zero at $z = 15/16$. LSFs are computed from tenth-order prediction coefficients. Figure 10 plots only the mean value (indicated by ▼ or Δ) and the spread of 99% of samples (indicated by a bar), rather than the probability density function to show the major features of the frequency distribution succinctly.



(a) From 54 male speakers, 2 sentences each

(b) From 12 female speakers, 2 sentences each

Fig. 10. — Distribution of LSFs derived from tenth-order LPC. For each LSF, the mean value and the standard deviation are shown for both voiced and unvoiced frames. LSFs for unvoiced frames are consistently higher than those for voiced frames. Note also that the mean values for the first two LSFs are closer together than the other adjacent LSFs.

According to Fig. 10, there is no significant difference between male and female voices for the sixth through the tenth LSFs (i.e., for frequencies above approximately 2 kHz). Even for the first through the fifth LSFs, the difference is primarily in the spread, not the mean value. In general, the frequency spread is greater for male voices, particularly for voiced F_3 and F_4, and unvoiced F_5. It is interesting to note that the unvoiced male and female speech is virtually identical in terms of the LSF means (Table 3). Implication of this is that the unvoiced speech segment does not contain much cue information which is related to speaker identification. Note also that the mean values lie very close to the equally spaced LSFs which have a flat spectrum.

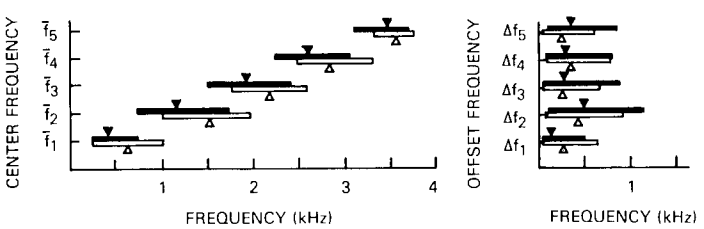
As an alternate representation of Fig. 10, Fig. 11 plots the center and offset frequencies of LSF pairs which are defined respectively as

$$\bar{f}_k = \frac{1}{2}(f'_k + f_k), \tag{56a}$$

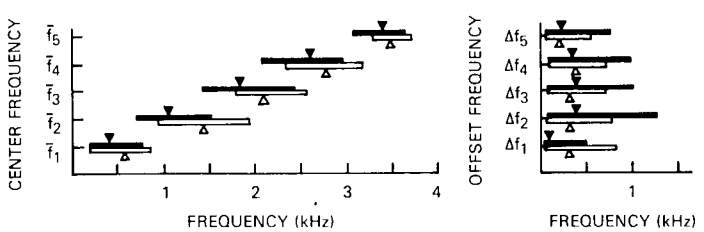
$$\Delta f_k = \frac{1}{2}(f'_k - f_k), \tag{56b}$$

Table 3 — Comparison of Equally Spaced Frequencies with LSFs Generated from Unvoiced Speech. These data are obtained from Fig. 10. Note that the mean values of the LSFs for male and female unvoiced speech are virtually identical.

Frequency Index	Equally Spaced Frequencies (Hz)	Unvoiced-Male Speech			Unvoiced-Female Speech		
		Mean LSF (Hz)	Difference		Mean LSF (Hz)	Difference	
			(Hz)	(%)		(Hz)	(%)
1	363.64	360	−4	−1	360	−4	−1
2	727.27	720	−7	−1	730	+3	+0
3	1090.91	1180	−11	−1	1170	−21	−2
4	1454.54	1550	+95	+7	1560	+105	+7
5	1818.18	1920	+102	+6	1920	+102	+6
6	2181.81	2240	+58	+3	2230	+48	+2
7	2445.45	2590	+145	+6	2590	+145	+6
8	2909.09	2950	+41	+1	2950	+41	+1
9	3272.72	3340	+67	+2	3340	+67	+2
10	3636.36	3640	+4	+0	3640	+4	+0



(a) From 54 male speakers, 2 sentences each



(b) From 54 female speakers, 2 sentences each

Fig. 11 — Distribution of the center and offset frequencies of LSF pairs. This figure is an alternate representation of Fig. 10. LSF pairs are made of the first and second frequencies, the third and fourth frequencies, etc. Note that the first offset frequency is much smaller than the other offset frequencies.

where f'_k and f_k are the kth LSFs which are the null frequencies of the sum and difference filters (see Fig. 7 or 8), and f'_k is numerically larger than f_k as listed in Table 2. There are two striking features of LSFs exhibited in Fig. 11. If speech is voiced, the offset frequency of the first LSF pair is much smaller than the others. This phenomenon is also shown in the LSF trajectories shown in Fig. 9. On the other hand, if speech is unvoiced, center frequencies are greater than those for voiced speech. Exploitation of these two features can eliminate grotesque voicing errors in the pitch-excited narrowband voice processor.

SPECTRAL SENSITIVITY OF LSFs

The spectral sensitivity of an LSF is determined by perturbing the LSF and finding the resulting change in the log spectrum of an all-pole filter. Error produced by quantization of the LSFs will be magnified by spectral sensitivity and appear as spectral error in the synthesized speech. We will show that not all LSFs are equally sensitive. Thus, to best use available bits during speech encoding, spectral-sensitivity analysis of the filter parameters is essential.

Observed Characteristics

Spectral distortions caused by LSF errors are considerably different from those created by reflection coefficients. Mentioning some of these differences is highly instructive.

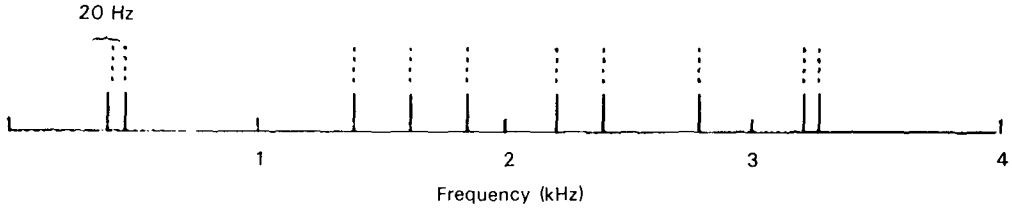
1. *Cross-coupling* — According to a previous investigation [16], spectral sensitivity in terms of reflection coefficient is expressed simply as the logarithm of the bilinearly transformed-reflection coefficients

$$g_i = \frac{1 + k_i}{1 - k_i}, \qquad i = 1,2,\ldots, \tag{57}$$

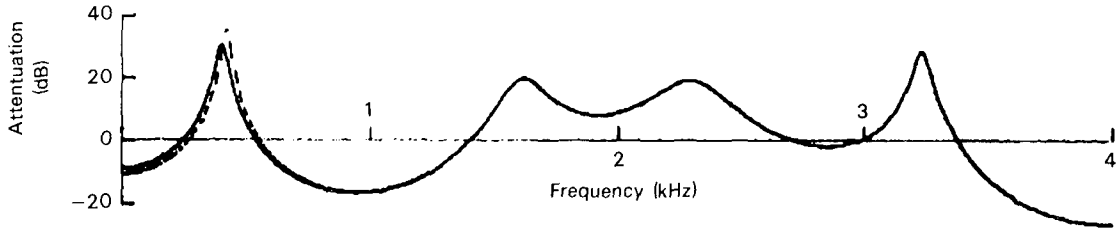
where k_i is the ith reflection coefficient. This is a most remarkable fact because the spectral sensitivity of a reflection coefficient is independent of other coefficients. As evident in Figs. 12 through 14, such is not the case with spectral error caused by an error in an LSF. Spectral error is not only dependent on the error magnitude of one particular LSF, but it is also dependent on the frequency separations with others. If an LSF which is removed from other LSFs has an error, its spectral sensitivity is small (Fig. 14). On the other hand, if the LSF with an error is in the proximity of other LSFs, then its spectral sensitivity is large (Figs. 12 and 13). An LSF may have an error in the low-frequency region (Fig. 12) or in the high-frequency region (Fig. 13), but the magnitudes of both spectral errors are large because other LSFs are in the vicinity.

2. *Localized spectral error* — Error in a reflection coefficient produces a spectral error in the entire passband. In contrast, an error in an LSF produces spectral error in the neighborhood of that particular frequency (Figs. 12 through 14). This phenomenon is somewhat similar to that of the channel vocoder or formant vocoder.

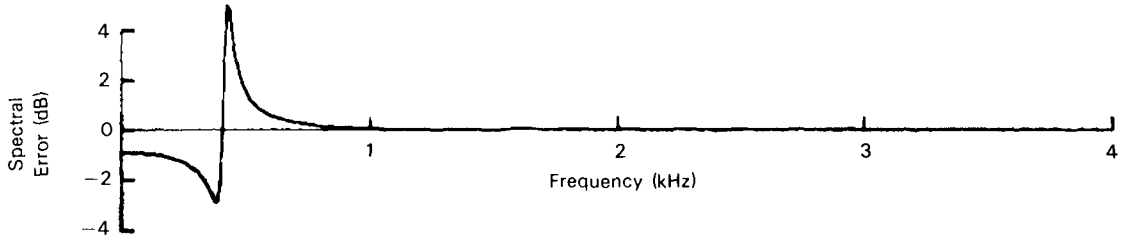
3. *Most critical filter parameters* — When reflection coefficients are used as filter parameters as in the current narrowband LPC, the first four coefficients are on the average the most critical parameters to the spectrum because these coefficient values are usually large for most vowels. When these four coefficients are error protected in the narrowband LPC or its modulator/demodulator, synthesized-speech quality is satisfactory even with a 5% random-bit error (viz., DoD Standard Narrowband LPC, and HF mode of Advanced Narrowband Digital Voice Terminal) [17]. On the other hand, when the LSFs are used as filter parameters, the first two frequencies are on the average the most critical parameters to the spectrum because their frequency separation is small for most vowels (Fig. 9). Table 4 lists numerically derived, spectral-sensitivity coefficients from voiced frames for both male and female voices. The first two LSFs are indeed the most sensitive.



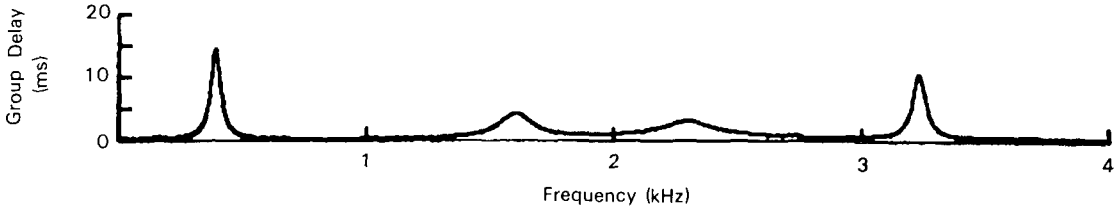
(a) Two sets of LSFs: one without error (solid lines), and the other with a 20 Hz error in the first LSF (dotted lines)



(b) LPC-analysis-filter-amplitude responses by the two sets of LSFs shown in (a)

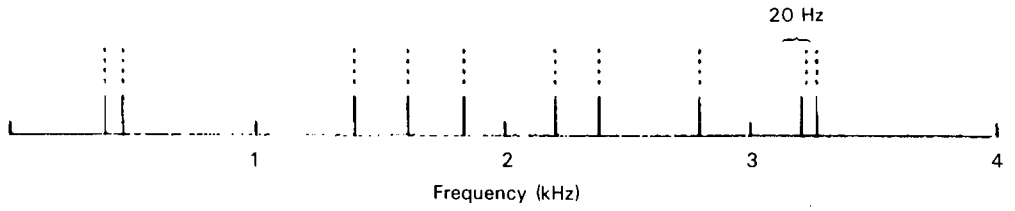


(c) Difference between the two amplitude responses shown in (b)

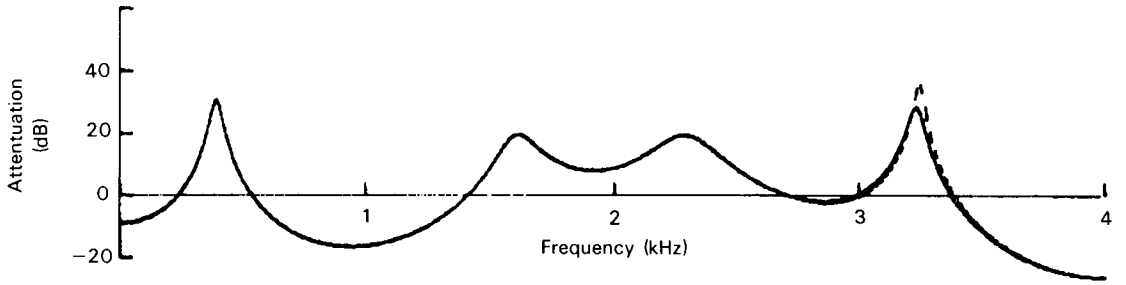


(d) Group delay of ratio filter

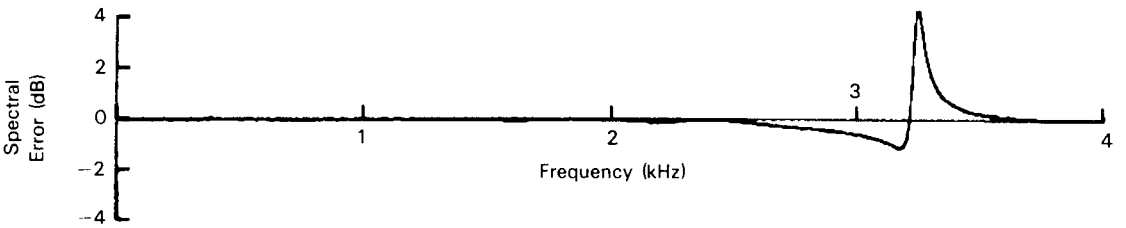
Fig. 12 — Spectral error caused by the error in a single LSF (Example #1). LSFs are obtained from tenth-order LPC analysis with preemphasis by a single-zero filter with zero at $z = 15/16$. The input speech is /i/ in "is." The spectral error is created by perturbing the first frequency by 20 Hz as indicated in Fig. 12(a). Note that the spectral error is concentrated near the perturbed LSF. The peak error is 5.03 dB, whereas the root-mean-square (rms) error over the entire passband is 0.71 dB (which is not a good indicator for the peaky error). Note that the group delay of the ratio filter at the perturbed frequency shown in Fig. 12(d) is rather large.



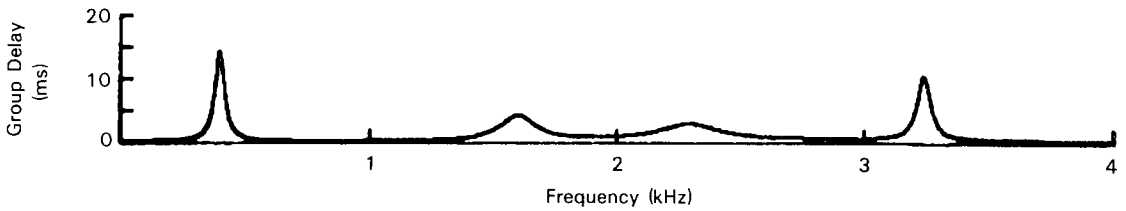
(a) Two sets of LSFs: one without error (solid lines), and the other with a 20 Hz error in the ninth LSF (dotted lines)



(b) LPC-analysis-filter-amplitude responses by the two sets of LSFs shown in (a)

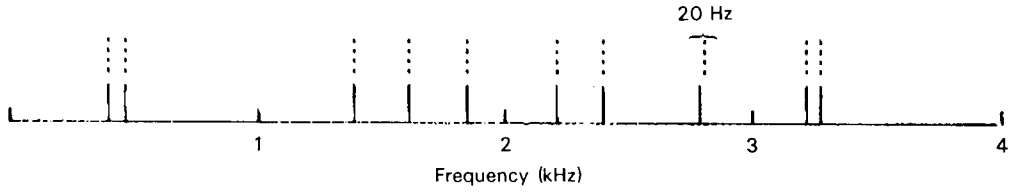


(c) Difference between the two amplitude responses shown in (b)

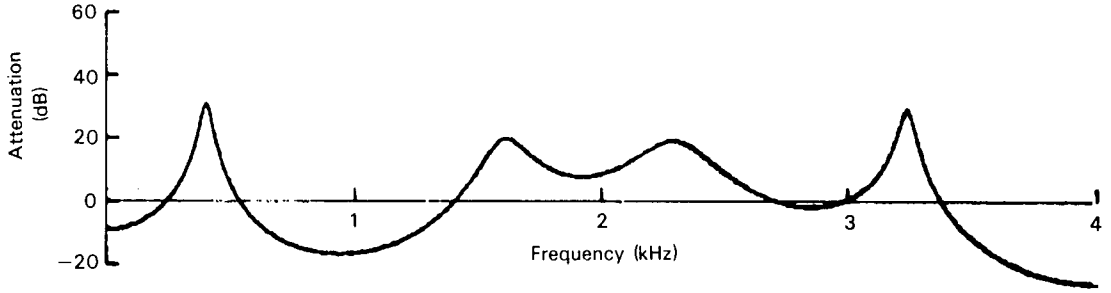


(d) Group delay of ratio filter

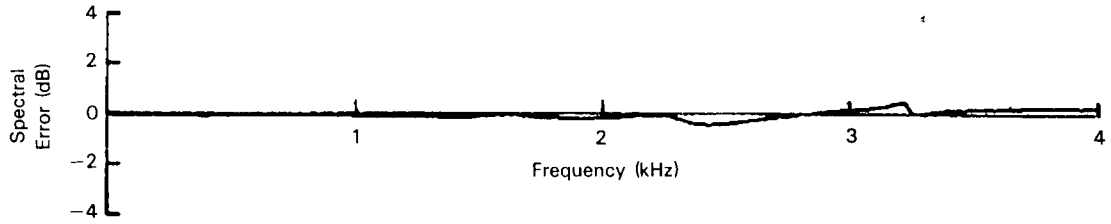
Fig. 13 — Spectral error caused by an error in a single LSF (Example #2). This example is identical to the preceding example except that the ninth LSF is perturbed by 20 Hz, rather than the first LSF. As shown in Fig. 13(d), the group delay of the ratio filter at the perturbed frequency is nearly as large as that of the first LSF. Thus, spectral error of a similar magnitude as the preceding example is expected. Computations show that the peak error is 4.28 dB and the rms error is 0.53 dB. Note that front vowels (such as /i/ shown in this figure) have strong upper resonant frequencies, particularly with preemphasis. Hence the spectral sensitivities of higher order LSFs are comparable to those of the first two LSFs. Since front vowels occur less frequently than do back vowels, the spectral sensitivity based on an ensemble of many samples (Table 4) does not indicate the phenomenon shown in this figure.



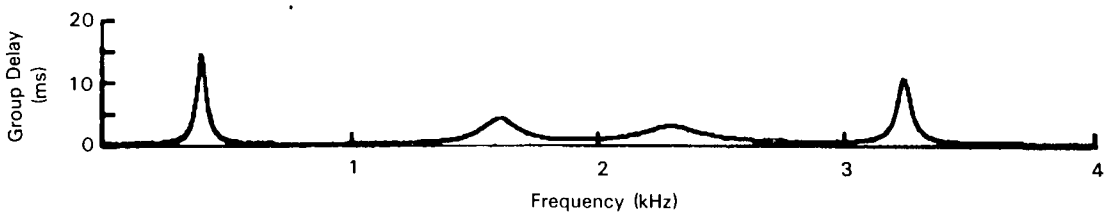
(a) Two sets of LSFs: one without error (solid lines), and the other with a 20 Hz error in the eighth LSF (dotted lines)



(b) LPC-analysis-filter-amplitude responses by the two sets of LSFs shown in (a)



(c) Difference between the two amplitude responses shown in (b)



(d) Group delay of ratio filter

Fig. 14 — Spectral error caused by an error in a single LSF (Example #3). This example is similar to the preceding two examples, except that the perturbed eighth LSF is rather isolated from the other frequencies. Thus, group delay of the ratio filter at that frequency is rather small as shown by Fig. 14(d). Computations indicate that the peak error is only 0.49 dB, and the rms error is 0.17 dB.

Table 4 — Spectral Sensitivities of LSFs from Voiced Speech. As before, LSFs are obtained from tenth-order LPC analysis with preemphasis by a single-zero filter with zero at $z = 15/16$. The figures for male voices are obtained with 4100 voiced frames from 25 different speakers, whereas the figures for female voices are obtained with 1972 voiced frames from 12 different speakers. Actually, the mean values converged to what they are shown in Table 3 after averaging approximately 500 frames. Thereafter, they did not vary significantly. As shown, spectral sensitivities of the first two LSFs are somewhat greater than others. Spectral sensitivities for unvoiced frames are of no particular interest. Since unvoiced speech contains only weak resonant frequencies, spectral sensitivities of all LSFs are approximately equal, somewhere around 0.01 dB/Hz.

LSF Index	Male Voices		Female Voices	
	Mean (dB/Hz)	Standard Deviation (dB/Hz)	Mean (dB/Hz)	Standard Deviation (dB/Hz)
1	0.0228	0.0096	0.0174	0.0084
2	0.0184	0.0066	0.0196	0.0069
3	0.0111	0.0048	0.0114	0.0034
4	0.0153	0.0056	0.0115	0.0046
5	0.0141	0.0067	0.0151	0.0054
6	0.0134	0.0054	0.0133	0.0053
7	0.0164	0.0062	0.0133	0.0043
8	0.0106	0.0033	0.0129	0.0040
9	0.0140	0.0054	0.0111	0.0029
10	0.0152	0.0056	0.0136	0.0049

Parametric Representation of Spectral Sensitivities

The transfer function of the LPC-analysis filter, as obtained from Eqs. (3), (9), (10), (15), and (16), is

$$A_n(z) = \frac{1}{2}\left\{(1 - z^{-1})\prod_{k=1}^{n/2}[1 - 2\cos(2\pi f_k t_s)z^{-1} + z^{-2}] + (1 + z^{-1})\prod_{k=1}^{n/2}[1 - 2\cos(2\pi f'_k t_s)z^{-1} + z^{-2}]\right\}, \tag{58}$$

where f_k and f'_k are the kth LSF pair. Spectral sensitivity of the LSF is a change of $A_n(z)$ in decibels in terms of a change in an LSF. To derive such an expression is not only untractable, but also a coupling of all frequency errors into the overall spectral error makes the use of such an expression virtually useless.

An alternative way of expressing the spectral sensitivity is through the use of an auxiliary variable (i.e., parametric representation of the spectral sensitivity). We observe from Figs. 12 through 14 that

the spectral sensitivity of an LSF is directly dependent on the group delay of the ratio filter at that frequency. The group delay of the ratio filter, expressed by Eq. (48), is dependent on the locations of all the LSFs, but it is particularly sensitive to the frequency separation between adjacent LSFs. Thus, spectral sensitivity may be related to the group delay through a single curve, independent of the LSF index, or mutual frequency separations among the LSFs.

Figure 15 is a scatter plot of spectral errors in terms of group delays associated with perturbed LSFs. No distinction has been made for the order of LSF index, nature of voicing (voiced or unvoiced), or category of speakers (male or female). Every sample that went into Fig. 14 had a single discriminating criterion; namely, the group delay of the ratio filter associated with the perturbed frequency. Considering 17,881 samples went into this plot, the clustering characteristic is remarkable. The best fit spectral sensitivity curve is expressed as

$$E = 0.0096\sqrt{D} \quad \text{dB/Hz}, \tag{59}$$

where E is the spectral sensitivity, and D is the group delay in milliseconds. The rms error between the entire samples and the best fit curve is only 0.00138 dB/Hz.

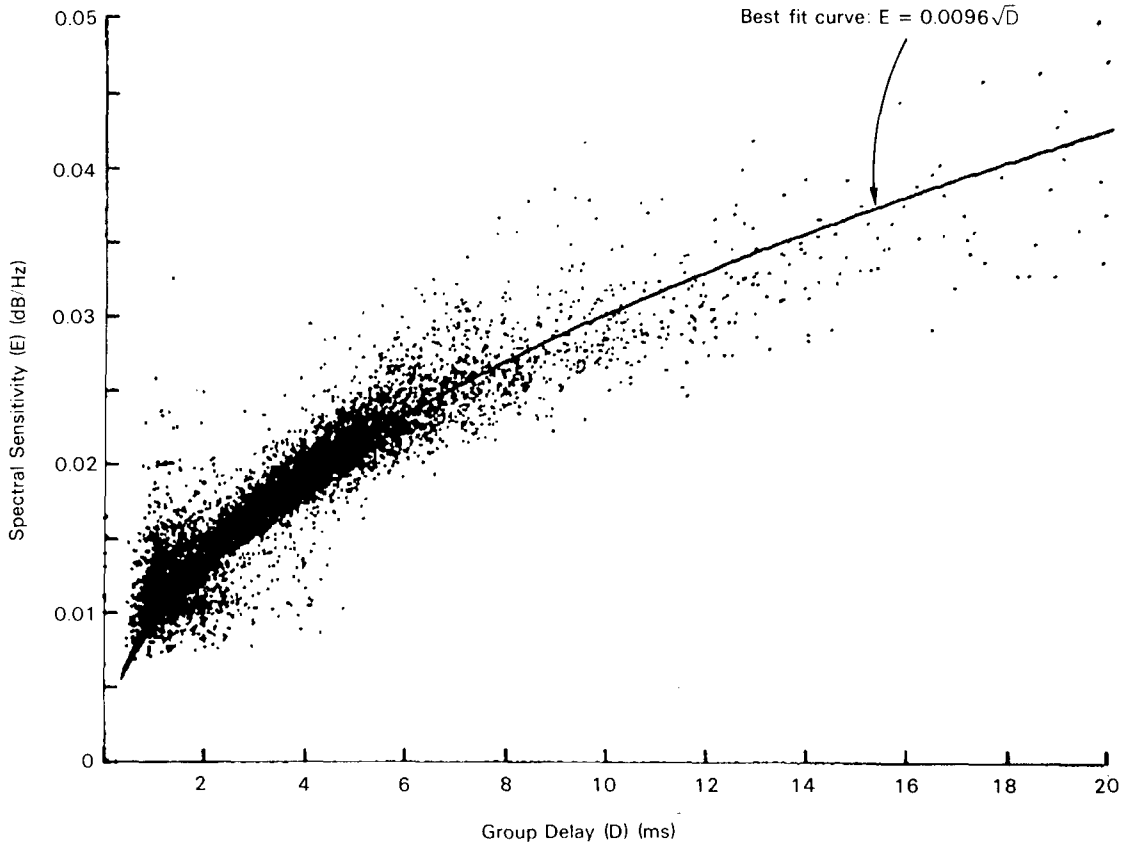


Fig. 15 — Scatter plot of spectral sensitivity in terms of group delay of the ratio filter. Each point in the plot represents the rms spectral error in decibels caused by the perturbation of any one of the first through tenth LSFs. The independent variable is the group delay of the ratio filter evaluated at the LSF. The single best fit curve represents the spectral sensitivities of the first through tenth LSFs. The total number of LSFs entered into this figure is 17,881. The distributions of samples are 62% for $D \leqslant 2$ ms, 86% for $D \leqslant 4$ ms, and 95% for $D \leqslant 6$ ms.

If LSFs are computed by the use of the ratio filter (i.e., Approach #2 in the previous discussion in this report), we can obtain the delays by simply taking the gradients of the phase angles at the crossing of $-\pi, -2\pi, -3\pi, \ldots$ radians where the LSFs are located. Thus, group delays are essentially the byproduct of LSF computations. (Even if only the LSFs are given, group delays can be computed by transforming the LSFs back to prediction coefficients, and recomputing the LSFs.)

PERCEPTUAL SENSITIVITIES

Inaccurate representation of LSFs caused by quantization errors results in spectral error in the synthesized speech. Spectral error, in turn, is perceived by the human ear. If the error is large, it leads to a misidentification of the phoneme. Inasmuch as the human ear makes the ultimate classification of sound, peculiarities of human auditory perception (in addition to spectral sensitivities of filter parameters discussed in the preceding section) must be fully exploited to encode speech efficiently.

Human perception of spectral distortion is a complex problem. Equal amounts of spectral distortion expressed in decibels can sound considerably different to the listener. For example, if the spectral distortion is stationary, the effect is much more bearable to the listener than the same amount of distortion that is time-variant. Likewise, high-frequency distortion is more tolerable to the listener than the same amount of distortion in the low-frequency region. Only recently has parameter quantization in voice processors been based on auditory perception, such as: masking effect, preference of the noise-spectral level below the speech spectrum (i.e., use of a noise shaper), and the like [18,19]. Only a few perceptual experiments are presented in this section; more experimentation in this area is highly desirable.

Perceptual Sensitivity to LSF Changes

It is well known that the amount of frequency variation in pitch that produces a perceived just-noticeable difference (JND) is approximately linear from 0.1 to 1 kHz, but the frequency variation needed to produce a JND increases logarithmically from 1 to 10 kHz [20]. We would like to perform a similar experiment with LSFs.

In essence, we use speech-like sounds rather than a single tone. We incrementally change one of the ten LSFs in order to create a set of closely related sounds. From these sounds, each listener decides his or her own JND in terms of the variation in one LSF. The JND is dependent not only on the magnitude of a frequency shift, but also the change of the spectral amplitude. Unfortunately, we do not have control over the spectral amplitude because it is implicitly determined by all the LSFs. As we have seen from Figs. 12 and 13, spectral change is particularly large near two closely spaced LSFs. To minimize the effect of the spectral-amplitude change on the determination of the JND, we chose equispaced initial LSF values.

The procedure used in this experiment is called the "method of limits" or "method of minimal change" [21]. An example of this method is the hearing test where the subject determines when he first hears a tone (as the level increases or ascends) and when it first becomes inaudible (as the level decreases or descends). The average of these two values gives rise to a threshold. After the subject listens to a number of the "ascending" and "descending" series of trials, the average of the thresholds determines an approximate sensitivity for the listener to that tone for the experimental conditions that were tested. By using a number of different subjects, a threshold is determined for the general population.

Because we would like to test the sensitivity of human perception to changes in LSFs, excitation parameters were fixed throughout the experiment with a voiced state having a pitch frequency of 100 Hz. The experiment is broken up into ten parts corresponding to the ten LSFs. Except for the LSF being perturbed, the remaining LSFs assume the equispaced values shown in Fig. 16. Part 1 of the

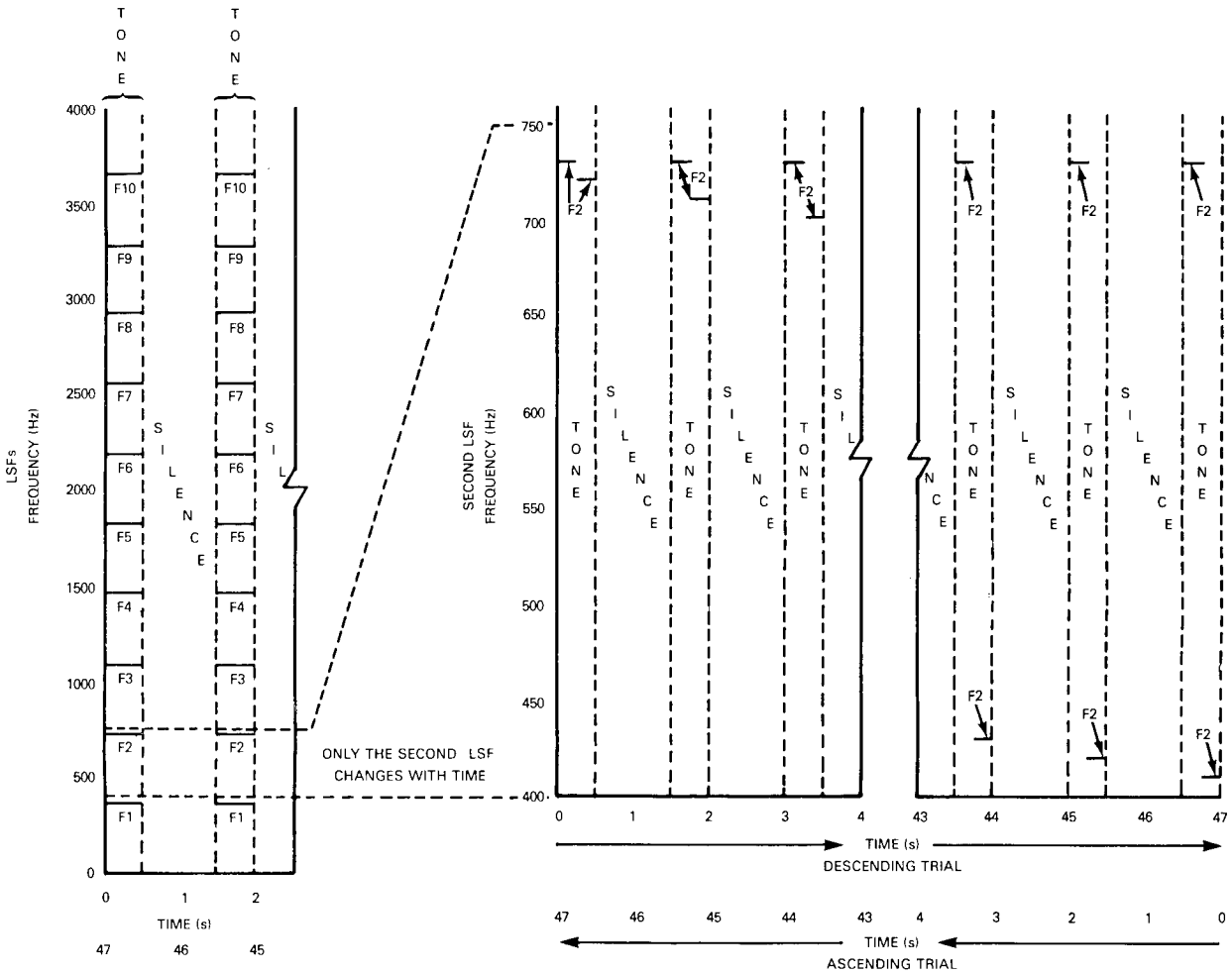


Fig. 16 — Part 2 of perceptual experiment showing second LSF changing with time

experiment deals with determining the JND in perception as the first LSF is perturbed while the remaining LSFs (two through ten) remain constant. Part 2 of this experiment allows the second LSF to be perturbed while the first and third-through-tenth LSFs remain fixed. The remaining parts to this experiment are organized in a like manner.

Since the procedure for finding the JND of a perturbed LSF is identical for each LSF, we will choose an LSF (namely, the second LSF) to describe the experiment (which we denote as part 2). In part 2, the first and third-through-tenth LSFs are fixed as shown in Fig. 16. The listener tries to determine the minimal reduction in frequency of the second LSF (descending case) required for a JND in perception. To determine this minimal amount of shift, the listener hears through a headset the output of a tape recorder having a series of computer-generated tones, each of which is one-half second in duration with one second of silence between tones. The first half of all the tones heard by the listener was generated with equispaced LSFs, but the last half of the tone (one-fourth second in duration) has the second LSF changing with each tone. In the decending case (from left to right as shown in Fig. 16), the listener determines at which point a JND is perceived. The ascending case (from right to left) is performed with the listener determining when the JND in the tones is no longer heard. The average of the two values obtained from the ascending and descending series of tones determine a threshold. The process just described is repeated to reduce the variance. The two thresholds are averaged to arrive at a threshold for a particular listener for the second LSF.

To arrive at statistics that represent a larger population, 16 listeners took the perception test. To minimize the effect of learning on the results, the order at which the parts of the experiment were played to the listeners varied. Also, half of the listeners heard the LSFs as they increased in index while the other half heard the LSFs as they decreased in index.

The minimum shift needed for detection is known as "absolute threshold" or *absolute limen* (Latin for threshold) [21]. The *absolute limen* is not unique as it will vary with such factors as testing procedures, number of listeners, and experimental setup.

The experimental results are plotted in Fig. 17. As noted, the JND in change of frequency is directly proportional to frequency. The difference in perceptual sensitivity of the first LSF at 364 Hz and that of the tenth LSF at 3.636 kHz is approximately two to one which is quite similar to that obtained by using a single tone [20] at these two frequencies.

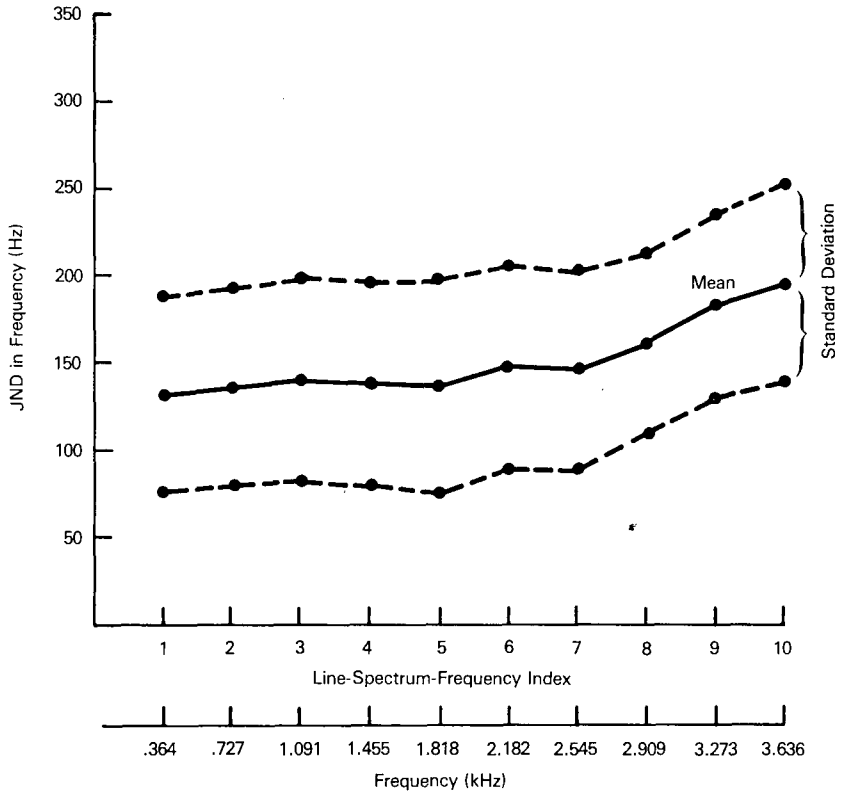


Fig. 17 — JND of perturbed LSFs

Sensitivity of DRT Scores to LSFs

The DRT is a perception test where the listener, on hearing a synthesized word, makes a choice between two rhyming words, one of which is correct. These two words differ by an initial consonant attribute. The DRT has been the most frequently used method for evaluating synthesized speech. All DoD-developed voice processors have been subjected extensively to the DRT. Initially, the DRT was used as a diagnostic means for improving the voice processor under development. Recently, however, it has also become a means for ranking voice processors.

Certain word pairs in the DRT list (there are 192 of them) are rather difficult to discriminate; for example, "fin" versus "thin," "sole" versus "thole," and "fad" versus "that." Even if we use unquantized filter parameters, synthesized speech is often not clear enough for a sure discrimination (conventional pitch-excitation signal used in the narrowband LPC is partly to blame for this). On the other hand, some other word pairs are not as difficult to discriminate, for example: "go" versus "Joe," "chair" versus "care," and "dint" versus "tint." For these words, we can quantize the filter parameters considerably more coarsely, and we can still distinguish one word of the pair from the other. In fact, choosing between "dint" and "tint" is more dependent on a sudden change in the loudness (i.e., the excitation signal parameters) than of the spectral content (i.e., filter parameters).

Because the sensitivity of DRT scores to LSFs is quite complex, we cannot express in simple terms the relationship between DRT scores and LSFs. We will, however, examine a few special cases which are useful for speech encoding.

Effect of Logarithmically Quantized LSFs on DRT Scores

The equitempered scale (or logarithmic scale) is a collection of frequencies in which any two adjacent frequencies form the same ratio. The equitempered scale is not only used for tuning a musical instrument, but it has been often used to quantize some voice processor parameters. For example, the fundamental pitch of the voiced-excitation signal of the narrowband LPC is quantized with an equitempered scale having 20 steps per octave (i.e., frequency resolution of 3.5%). Thus, encoding the pitch frequency range from 50 to 400 Hz requires only 60 discrete values (six bits for binary representation).

Likewise, formant frequencies of a formant vocoder may be encoded more efficiently if the log scale is used [22]. The JNDs of formant frequencies are on the order of 3 to 5% although this is greatly dependent on the proximity of the formants to one another. A frequency resolution of 3% is equivalent to that of the equitempered scale having 24 steps per octave. On the other hand, a frequency resolution of 5% is equivalent to having 15 steps per octave which is slightly coarser than the chromatic scale.

As noted above, frequencies may be quantized more efficiently without compromising the perceptual quality of the synthesized speech. Thus, we are interested in the effect of logarithmically quantized LSFs on the DRT score. To carry out this investigation, we originally chose four equitempered scales for frequency quantization: 24, 18, 15, and 12 steps per octave. The corresponding frequency resolutions are approximately 3%, 4%, 5%, and 6%.

First, we tested the case of a 6% frequency resolution (i.e., 12 steps per octave). For this case, frequencies from 400 to 3200 Hz (three octaves) are quantized to $3(12)+1 = 37$ values (including both end frequencies). Two additional frequencies below 400 Hz and three additional frequencies above 3200 Hz cover the entire frequency range of interest. Thus, the total number of frequencies is only 42. Hence, the number of combinations for choosing 10 frequencies out of 42 possible frequencies is 1,471,442,974; $\approx$ a binary-rated figure higher than this number is 2^{31}. Actually, the number of combinations is much less than this figure because certain combinations do not occur with speech (i.e., the tenth LSF would not likely be less than 2500 Hz, or the first LSF be above 1200 Hz, etc.).

As listed in Table 5, the resulting DRT score is only 1.8 points below that for using unquantized filter parameters, and virtually equal to that of the 2400-b/s LPC. Even if all LSFs are independently quantized to 33 bits per frame, the result is as good as that of the 2400-b/s LPC which uses 41 bits per frame for filter parameters. Since the drop in DRT score is so small with a 6% frequency resolution for LSFs, we did not run DRTs for 3%, 4%, and 5% frequency resolutions.

Table 5 — DRT Scores when LSFs Are Quantized Logarithmically. The bits referred to are the number of bits per frame used to encode filter parameters. For comparison, DRT scores with unquantized reflection coefficients and the 2400-b/s LPC are also listed. In all cases, the excitation signal is identical to that used in the 2400-b/s LPC. The scores are based on three male speakers ("LL," "CH," and "RH"). The use of LSFs saves at least 8 bits per frame for speech quality similar to 2400-b/s LPC.

Sound Classification	Reflection Coefficients		Line-Spectrum Frequencies	
	Unquantized	Quantized to 41 Bits[a]	Quantized to 31 Bits[b]	Quantized to 33 Bits[c]
Voicing	90.6	91.7	91.7	90.6
Nasality	95.3	93.7	90.4	93.2
Sustention	81.0	79.7	79.2	83.9
Sibilation	90.1	89.6	91.9	91.4
Graveness	87.0	80.7	80.2	75.5
Compactness	94.8	94.8	94.3	95.6
Total	89.8	88.4	88.0	88.4

[a] First-through-tenth reflection coefficients are independently quantized at 5,5,5,5,4,4,4,4,3, and 2 bits as used in the DoD Standard 2400-b/s LPC.
[b] Maximum number of LSF combinations inherent in a 6% frequency resolution is 31 bits (see text).
[c] Each LSF is not only quantized to have a 6% frequency resolution, but each is also bounded in range. First through tenth LSFs are represented by 4,4,4,4,4,3,3,3,2, and 2 bits.

Effect of LSF Elimination on the DRT Score

Most formant vocoders do not transmit any information on the fourth and fifth formant frequencies. One reason is that these formants are not always present in speech; therefore, their tracking is a major problem. But a more significant reason for not transmitting them is that speech intelligibility is adequate using information related to the first three formant frequencies.

We are interested in finding out whether the elimination of some of the higher indexed LSFs also results in a graceful degradation of speech intelligibility as in the formant vocoder. Higher indexed LSFs describe the speech-spectral envelope in the high-frequency region, similar to higher formant frequencies. We would like to know the DRT score sensitivity when the higher indexed LSFs are not present.

When the formant vocoder uses only the first three formant frequencies, the synthesized speech sounds somewhat dull because the speech bandwidth is only about 3 kHz (third formant frequency does not swing above 3 kHz too often). This is not necessarily the case for a voice processor employing LSFs as filter parameters because the receiver can regenerate speech having the full bandwidth by reintroducing eliminated LSFs. We can reintroduce eliminated LSFs because there are always a fixed number of LSFs in any speech at any time. Exact locations of these high-indexed LSFs are important, but they are not as critical as one might expect. Since all LSFs are naturally ordered, eliminated LSFs must be reintroduced somewhere between the highest LSF transmitted and the upper cutoff frequency (see typical LSF trajectories shown in Fig. 9).

We may place each omitted LSF at an equal spacing from the highest LSF transmitted to the upper cutoff frequency. The spacing will actually vary from frame to frame because the highest LSF transmitted will vary. The synthesized speech sounds similar to that using all ten original LSFs. A casual listener cannot discern the difference.

DRT scores, however, do show some differences. As listed in Table 6, when the highest LSFs (i.e., ninth and tenth LSFs) are eliminated there is a reduction of 3.3 points in the score. As two additional LSFs are eliminated there is an additional 1.5-point drop in the score.

Table 6 — Three Male DRT Scores in Terms of Number of LSFs Eliminated from the Highest Index. The eliminated LSFs are substituted with artificially derived values at the receiver. The excitation signal is identical to that used in the narrowband LPC. The most significant degradation occurs for the attribute "graveness," which tests "weed" versus "reed," "bid" versus "did," and "peek" versus "teak," among others. Discrimination of these word pairs requires accurate upper formant frequencies. The scores are based on three male speakers ("LL," "CH," "RH").

Sound Classification	Number of LSFs Eliminated		
	0	2	4
Voicing	90.6	89.6	90.1
Nasality	95.3	94.8	93.2
Sustention	81.0	74.7	77.6
Sibilation	90.1	90.1	83.3
Graveness	87.0	75.0	74.2
Compactness	94.8	94.5	91.7
Total	89.8	86.5	85.0

Instead of eliminating LSFs we can also eliminate some of the offset frequencies of LSF pairs defined by Eq. (56b). The eliminated offset frequencies may be reintroduced at the receiver based on their respective mean values as shown in Fig. 11. As noted from the DRT scores listed in Table 7, the two highest offset frequencies may be omitted without degrading speech intelligibility. Thus, a tenth-order LPC system can use only eight filter parameters.

Table 7 — One Male DRT Scores in Terms of the Number of Offset Frequencies of LSF Pairs Eliminated from the Highest Index. The eliminated offset frequencies are substituted with their statistics shown in Fig. 11. The excitation signal is identical to that used in the narrowband LPC. The scores show that the two highest offset frequencies can be eliminated from transmission without degrading intelligibility. The score is based on one male speaker ("RH").

Sound Classification	Number of Offset Frequencies Eliminated			
	0	1	2	3
Voicing	90.6	95.3	91.4	93.0
Nasality	97.7	96.9	95.3	93.0
Sustention	76.3	79.7	78.7	77.1
Sibilation	89.8	89.1	91.4	93.0
Graveness	83.3	76.1	90.3	74.0
Compactness	96.1	95.3	96.9	89.8
Total	89.0	88.7	89.0	86.7

IMPLEMENTATION OF AN 800- AND 4800-b/s VOICE PROCESSOR

We can use a single voice processor to generate these three rates because there are many computation commonalities. For example, all three rates require LPC analysis and synthesis. The 800-b/s mode, as in the 2400-b/s LPC, needs a pitch tracker and voicing decision. With all of these voice processing algorithms in one hardware unit, we can reduce the multiplicity of hardware and simplify logistics.

During the preamble period, one data rate is selected by the operator, and the choice of data rate is transmitted to the receiver as header data. Based on the received-header information, the operator selects one of the three voice processing algorithms stored in memory. To make implementation simpler, the speech waveform is sampled at a fixed rate of 8 kHz, and the frame size is fixed at 180 samples. Likewise the synchronization-bit pattern is identical to that used in the 2400-b/s LPC (namely, alternating "1" and "0" every 54 bits).

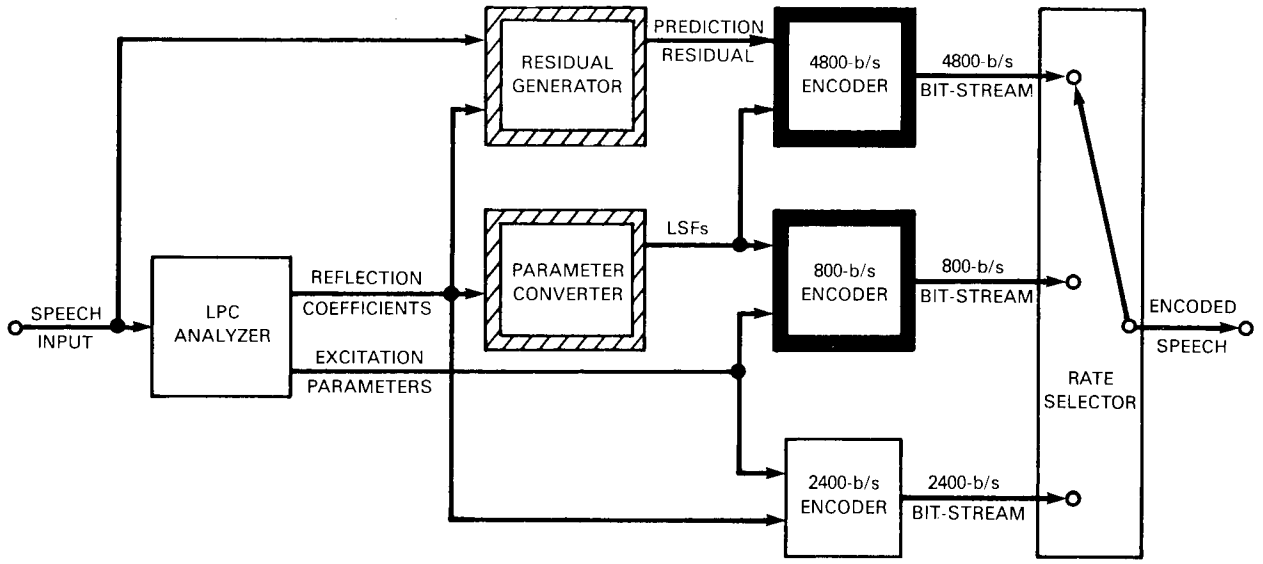
Figure 18 is a block diagram of the voice processor. As noted, there are many shared functional blocks among the different rates. The single-lined blocks are those used in the 2400-b/s LPC, and they do not need any further elaboration because they are well defined in Federal Standard 1015. The hatched blocks have been discussed earlier in this report. Thus, the present discussion concentrates on the heavy-lined blocks (encoders and decoders of the 800- and 4800-b/s voice processors). The 800-b/s voice processors may be named the "pitch-excited line-spectrum vocoder," and the 4800-b/s voice processor may be named the "nonpitch-excited line-spectrum vocoder."

800-b/s Encoder/Decoder

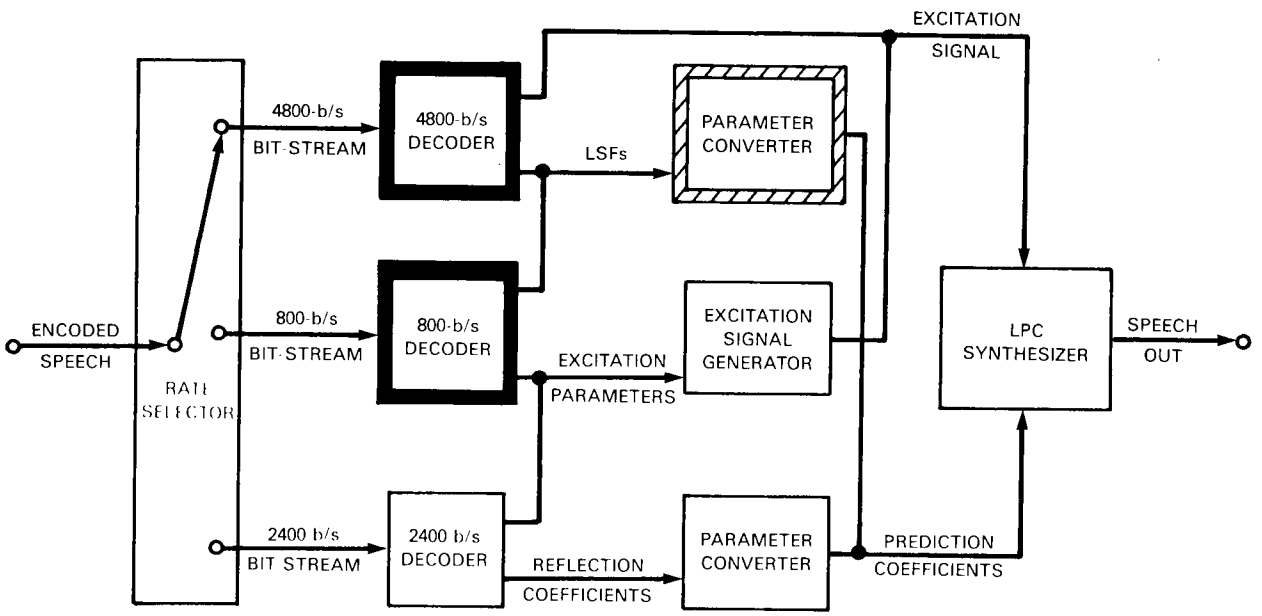
A data rate of 800 b/s is approximately 1% of the data rate of unprocessed digitized speech. When the data rate is compressed to this extreme limit, degradation of both speech intelligibility and quality is inevitable. We would like to review the scope of this severe compression of speech information, and then we will proceed carefully with the specification of an 800-b/s encoder/decoder.

Speech can be generated at an average rate of 100 b/s as demonstrated by the VOTRAX speech synthesizer [23]. Since the VOTRAX generates speech by a set of rules, the resulting synthesized speech does not imitate any one particular speaker. If a voice processor is designed to imitate an actual person's voice, the required data rate increases dramatically. For example, "Speak and Spell" (devised by Texas Instruments (TI)) is a speech synthesizer that imitates a speaker's voice. TI analyzed one person's voice (a good broadcasting voice from an announcer at a local radio station). To generate speech data, TI segmented speech visually using a sophisticated interactive-computer system. The speech data from each segment was repeatedly played back for evaluation. If needed, speech data from one segment was replaced by other speech data of similar sounds in order to achieve better speech quality. The resulting speech data was stored in "Speak and Spell" for synthesis. Even under such an ideal analysis/synthesis condition and using only one speaker, the data rate of "Speak and Spell" varies from 600 to 2400 b/s [24] with unknown distribution. If it is symmetrically distributed about its mean, the average data rate is 1500 b/s. We do not know what the data rate would be in "Speak and Spell" if the number of speakers is increased, but we do know that even 2400 b/s is not sufficient for effortless communications (see Fig. 1). This is somewhat disturbing since one out of 2^{41} (2.2 trillion) spectral sets are transmitted when speech is voiced. As will be shown, these 2.2 trillion spectral sets are reduced to 3840 in the 800-b/s voice processor. This 500,000,000-to-1 reduction of spectral information introduces significant speech degradation, particularly in a multispeaker environment with casual conversational speech.

Speech degradation occurs not only in conversational tests, but also in the DRT scores as well. One of the six attributes of the DRT most sensitive to the size of available spectral sets is "graveness" which tests the listener's discrimination to such words as: "did" versus "bid" and "weed" versus "reed."



(a) Transmitter



(b) Receiver

Fig. 18 — Block diagram of three-rate processor. Single-lined blocks are those used in DoD-standard 2400-b/s LPC which will not be discussed. The hatched blocks have been explained earlier in this report. Heavy-lined blocks are explained in this section.

The main spectral difference between these word pairs is in the second and third formant trajectories [25]; this difference becomes ambiguous as the number of available spectral sets decreases. The score for "graveness" is typically in the low 90s for 16,000-b/s voice processors and between the upper 70s to the lower 80s for the 2400-b/s LPCs. The variability of the score is greater at the 2400-b/s data rate. Even a slightly increased hum level in the front-end analog circuit can lower the score for "graveness" significantly. According to available test data, the score for "graveness" at 800 b/s is further down to somewhere between the upper 50s and the lower 70s. We note that a higher score for "graveness" results in a higher overall DRT score because the scores for the other attributes do not degrade as significantly with a reduction in the data rate. This points out the importance of the size of the available spectral sets.

The overall speech intelligibility depends on many interrelated factors: choice of filter parameters, method of quantization, number of available spectral sets (which is dependent on how pitch and amplitude information are quantized), partition of voiced and unvoiced spectral sets, and most important, exploitation of the spectral sensitivity of the speech synthesis filter and auditory perception characteristics of the human ear. We will investigate all these areas.

Pitch Encoder/Decoder

To make more bits available for encoding filter parameters we will encode the pitch period as coarse as the ear can tolerate. But we do not contemplate using artificial or constant pitch because the use of natural pitch is essential for making synthesized speech more acceptable to listeners. The pitch value, however, need not be exact because there are many acceptable pitch contours for a given speech. According to an experiment conducted by Gold and Tierney with an 8-kb/s channel vocoder, the pitch contour may be lowered as much as 3% or raised as much as 2% without being sensed by the listener [26]. The listener in a two-way communication link will not know the actual pitch contour because there is no way for making comparison between the two versions. Therefore, the pitch contour can be considerably off from the actual pitch contour.

It is significant to note that pitch has little influence on the DRT score. In fact, use of constant pitch produces as good an overall DRT score as the use of natural pitch although it generates mechanical sounding speech. Table 8 shows a comparison of DRT scores we recently obtained.

Table 8 — Effect of Constant Pitch on DRT Score for Three Males. Pitch does not carry any information related to initial consonant. Therefore, DRT scores are not affected by the use of constant pitch in the 2400-/bs LPC.

Sound Classification	2400-b/s LPC	
	Normal Pitch	Constant Pitch
Voicing	90.6	96.4
Nasality	95.3	98.4
Sustention	81.0	84.4
Sibilation	90.1	88.0
Graveness	87.0	77.6
Compactness	94.8	95.3
Total	89.8	89.4

Because of the low listener acceptance of constant pitch, we transmit natural pitch, but only once per three frames. This is permissible because the rate of change of pitch is not as great as other parameters. For a 2400-b/s LPC, the pitch resolution is 3.5% (i.e., 20 steps per octave). We will quantize pitch into 5 bits, approximately 12 steps per octave from 67 to 400 Hz (Table 9). This pitch resolution is somewhat coarser than what we desire, but this is a tradeoff made to provide more filter-parameter sets for the 800-b/s voice processor. We do not differentially encode pitch so that the voice algorithm responds quickly to transitions from male to female voices, and vice versa.

Table 9 — Quantized Pitch Values. For a speech-sampling frequency of 8 kHz, a pitch period of 20 samples corresponds to a fundamental pitch frequency of 400 Hz. On the other hand, a pitch period of 120 samples corresponds to a pitch frequency of 66.667 Hz. As noted, pitch is approximately quantized logarithmically at 12 steps per octave (similar to the chromatic scale).

Pitch Code	Pitch Period	Pitch Code	Pitch Period	Pitch Code	Pitch Period
0	20	12	40	24	80
1	21	13	42	25	85
2	22	14	44	26	90
3	23	15	47	27	95
4	24	16	50	28	101
5	26	17	53	29	107
6	28	18	57	30	113
7	30	19	60	31	120
8	32	20	63		
9	34	21	67		
10	36	22	71		
11	38	23	75		

Amplitude Information Encoder/Decoder

In addition to the pitch period discussed above, amplitude information is another nonfilter parameter whose resolution may be made coarser to allow for more bits to encode the filter parameters. Amplitude information is the rms value of preemphasized speech for each frame. It controls the loudness of the synthesized speech. We will encode amplitude information into one of 16 3-dB steps. In comparison with a 2400-b/s LPC, it is 1 bit less. To best use the available 4 bits, we will use an automatic gain control at the front end of the voice processor. Furthermore, we will multiply the rms value of unvoiced speech by a factor of two prior to quantization because it is naturally lower in comparison to that of voiced speech, then divide by a factor of two after gain calibration.

Like the pitch period encoding, we will not differentially encode amplitude information. According to our experience, differentially encoded amplitude information results in a noticeable reduction in "sustention," one of the six attributes of the DRT. "Sustention" tests the discriminability between "box" versus "vox," and "thick" versus "tick," among others.

Bit Allocation

After bits are allocated for pitch and amplitude information, the remaining bits are assigned to filter parameters. Since pitch is transmitted once for every three frames, it is convenient to group three frames together although amplitude information and filter parameters are transmitted once per frame.

Since three frames are grouped into one, only 1 synchronization bit is needed for three frames. The total number of bits per three frames at 800-b/s is 54 for a frame rate of 44.44 Hz (as used in the DoD-narrowband LPC). Table 10 lists bits allocated for each parameter. Because of the bits saved during the encoding of the pitch period and the amplitude information, the total number of allowable spectral sets is as much as 2^{12} (i.e., 4096). This figure is two to four times greater than that used in some other 800-b/s voice processors.

Table 10 — Bit Allocation for Each Parameter

Parameter	Bits
Synchronization	1
Pitch Period	5
Amplitude Information	12[a]
Filter (with voicing decision)	36[b]
Total	54

[a] Derived from 4 bits each from three frames
[b] Derived from 12 bits each from three frames

Although the voicing decision is often encoded with a separate bit, the voiced and unvoiced segments are neither equally probable nor equally significant. Hence, we resort to the voicing decision being implicitly included in the filter parameter information. As we discuss in the next section, 4096 spectral sets are partitioned into voiced and unvoiced spectral sets. When a voiced spectrum is transmitted, the pitch excitation is used for speech synthesis. On the other hand, when an unvoiced spectrum is transmitted, random noise excitation is used for speech synthesis.

Partition of Voiced and Unvoiced Spectral Sets

The allowable 4096 spectral sets are partitioned into two disjoint sets: one for voiced speech and the other for unvoiced speech. We need fewer unvoiced spectral sets because an unvoiced spectrum need not be represented precisely. We are accustomed to hearing a wide range of fricative spectral variations from person to person [27]. Furthermore, we identify some fricative sounds under the influence of formant transitions in the neighboring vocalistic segments [28]. Hence, there is a many-to-one transform between the unvoiced speech spectrum and its perception to the human ear. The 2400-b/s LPC exploits this phenomenon by having a ratio of allowable voiced spectral sets to unvoiced spectral sets of 2^{41} to 2^{20} (i.e., 2.2 trillion to 1 million).

When 4096 spectral sets are available in an 800-b/s voice processor, the number of unvoiced spectral sets may be somewhere around 256 (therefore, the number of voiced spectral sets is 3840). This figure is based on our experimentation with a previous 800-b/s voice processor [29] which quantized the reflection coefficients vectorially using the quadratic difference of the log-area-ratio as the distance measure. According to our subsequent experimentation with LSFs as filter parameters, we have not found any reason for changing this partition.

LSF Encoder/Decoder

An ideal filter-parameter encoder encodes all the perceptually indistinguishable spectral sets into one of the codes; yet, none of the codes represents the spectrum of sounds unrelated to speech (such as the crow of a rooster). Since each filter parameter set represents one distinct sound, an ideal filter-parameter encoder has to be in the form of a block encoder (i.e., a vector quantizer) with distinct

sound spectra in the memory. The encoder compares the given speech spectrum with the stored spectral sets, and transmits the index of the nearest neighbor based on the chosen distance measure. The decoder reads out a set of filter parameters that corresponds to the received spectral index.

In practice, however, we will never be able to design such an ideal encoder as described above because: (a) we do not have all the representative speech samples (males and females, young and old, northerners and southerners, and normal and abnormal voices) from which allowable spectral sets are extracted; (b) even if we have them, we cannot possibly sort out all the bogus spectra generated by a mixture of speech sounds within the analysis window (this happens often at speech transitions).

Nevertheless, with some simplifications, approximations, and assumptions, usable vector quantizers operating at a data rate between 600 and 800 b/s have been devised in recent years. Some are based on the channel vocoder [30,31], others are based on the spectral-envelope-estimation vocoder [32], or the LPC [29,33]. Speech intelligibility, according to published accounts, varies from unit to unit. But, in general, speech intelligibility at 800 b/s is about five to ten points lower than that of a similar device operating at 2400 b/s. Thus, there is an appreciable amount of speech degradation from 2400 to 800 b/s.

In this report, we introduce another 800-b/s vector quantizer which is based on what we call the "line-spectrum vocoder." In one sense, this is an LPC because filter parameters are nothing more than transformed prediction coefficients. In another sense, this is somewhat akin to the channel vocoder, formant vocoder, or spectral-envelope-estimation vocoder because it uses frequencies as filter parameters, and a change in a frequency results in a spectral change primarily near that frequency. Thus, the line-spectrum vocoder combines a good spectral peak representation capability of the LPC and the frequency-selective-quantization property of the channel, formant, and spectral-envelope-estimation vocoders.

Furthermore, our distance measure for selecting a nearest neighbor spectral set is not only based on the spectral sensitivity of the individual LSF, but it is also based on the hearing sensitivity of the human ear. Inclusion of hearing sensitivity into the distance measure makes a great deal of sense for our vocoder application because the human ear makes the ultimate evaluation of speech quality. Perceptually motivated distance measures have been employed previously by Gold for the channel vocoder [31], and Paul for the spectral-envelope-estimation vocoder [32].

Distance Measure

Our distance measure is expressed as the rms of the weighted LSF differences between two sets of LSF vectors; namely, $\{F_a\}$ and $\{F_b\}$ with each comprised of ten LSF components. Thus,

$$d(F_a,F_b) = \sqrt{\frac{1}{10}\sum_{i=1}^{10}\{w(i)[F_a(i) - F_b(i)]\}^2},$$
$$= d(F_b,F_a), \tag{60}$$

where $w(i)$ is the i th weighting coefficient that transforms the LSF difference to a spectral difference which is more meaningful to our auditory perception. The weighting coefficient, $w(i)$, is a nonnegative number that is normalized for convenience to a value between 0 and 1.

If we are concerned only with spectral distortions, the weighting factor should be proportional only to the spectral-sensitivity coefficient of the individual LSF (Fig. 15) because it converts LSF differences to spectral differences. The distance measure, normalized to have a value between 0 and 1, would be

$$w(i) = \frac{0.0096\sqrt{D(i)}}{0.0096\sqrt{D_{\max}}} = \sqrt{\frac{D(i)}{D_{\max}}}, \tag{61}$$

where $D(i)$ is the group delay (in milliseconds) of the ratio filter associated with the ith LSF for either F_a or F_b, whichever is largest, and D_{max} is the maximum group delay observed from speech (i.e., approximately 20 ms as shown in Fig. 15).

The weighting coefficient expressed by Eq. (61) does not account for our peculiar hearing sensitivity. We know that spectral distortions in the spectral valleys are perceptually less significant than those near the spectral peaks. According to Flanagan [22], intensity limens for harmonic components located in the spectral valleys can be quite large, as much as +13 dB to $-\infty$ dB. We can incorporate this kind of hearing insensitivity to the weighting coefficient by use of the spectral-sensitivity curve (Fig. 15) which has been desensitized for smaller group delays. As shown in Figs. 8 and 12, the LSFs in the spectral valleys are associated with smaller group delays.

We would like to discuss the values of smaller group delays. As we recall, when the input signal has a flat spectrum without resonant peaks or nulls, all LSFs are equally spaced. The group delay of the ratio filter at any LSF (in fact, anywhere in the passband for this particular case) equals 11/8000 s or 1.375 ms, assuming a tenth-order LPC and a 4 kHz upper cutoff frequency. Smaller group delays are referred to as group delays less than 1.375 ms, and they are associated with LSFs mainly in the spectral valleys.

We would like to lower the spectral-sensitivity curve for smaller group delays. A simple and satisfactory solution is to modify the original spectral-sensitivity curve by a ramp function for smaller group delays where the ramp function passes through the origin and the original spectral-sensitivity curve at $D = 1.375$ ms. Since the ramp function is inscribed below the original spectral-sensitivity curve, the LSF difference is less sensitive to the spectral error for smaller group delays. Referring to Eq. (59), the original spectral-sensitivity curve is

$$E = 0.0096\sqrt{D} \qquad \text{for } 0 \leqslant D \leqslant D_{max}, \tag{59}$$

which is modified to

$$E = \begin{cases} 0.0096\sqrt{D} & \text{for } 1.375 \leqslant D \leqslant D_{max} \\ \dfrac{0.0096}{\sqrt{1.375}}\, D & \text{for } D < 1.375 \end{cases} \tag{62}$$

Thus, the weighting factor for the distance measure which includes both the spectral sensitivity of the individual LSF and the hearing sensitivity to spectral distortions near the spectral valleys, as obtained by the use of Eq. (62), is

$$w(i) = \begin{cases} \sqrt{\dfrac{D(i)}{D_{max}}} & \text{for } 1.375 \leqslant D(i) \leqslant D_{max} \\ \dfrac{1}{\sqrt{1.375\, D_{max}}}\, D(i) & \text{for } D(i) < 1.375 \end{cases}, \tag{63}$$

where $D(i)$ is the group delay associated with the i th LSF of either $\{F_a\}$ or $\{F_b\}$, whichever is larger.

Although the weighting factor expressed by Eq. (63) is preferred to that expressed by Eq. (61) for vocoder application, there is still another factor that must be incorporated in the weighting function;

namely, a gradual loss of our hearing resolution with increase in frequency. Thus, a more complete weighting function modified from Eq. (63) is in the form of

$$w(i) = \begin{cases} u(f_i)\sqrt{\dfrac{D(i)}{D_{\max}}} & \text{for } 1.375 \leqslant D(i) \leqslant D_{\max} \\ u(f_i)\dfrac{1}{\sqrt{1.375\ D_{\max}}}\ D(i) & \text{for } D(i) < 1.375 \end{cases}, \tag{64}$$

where $u(f_i)$ is the relative sensitivity of our hearing to frequency difference which depends on the nature of the tone (Fig. 19). A good approximation to the relative hearing sensitivity to frequency difference is

$$u(f_i) = \begin{cases} 1 & \text{for } f_i < 1000Hz \\ \dfrac{-0.5}{3000}\ (f_i - 1000) + 1 & \text{for } 1000 \leqslant f_i \leqslant 4000Hz \end{cases} \tag{65}$$

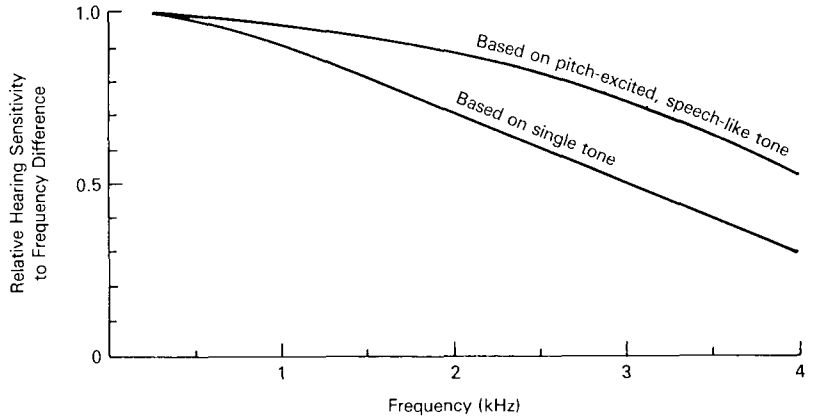


Fig. 19 — Relative hearing sensitivity to frequency differences. This figure shows our hearing sensitivity to discriminating frequency difference as a function of frequency. One curve is based on the JND of a single tone [20], and the other curve is based on the JND of pitch-excited, speech-like sound with a relative flat spectral envelope in which one out of ten LSFs is perturbed (Fig. 16). We expect the relative hearing sensitivity curve of speech sounds to be located somewhere between these two curves.

Template Collection

As required by any pattern-matching or recognition process, we need to form a template collection that partitions the available LSF sets into a set of clusters. One of the most frequently used cluster-analysis methods is the c-mean algorithm [34]. This method is executed in four steps:

Step 1 (Initialization): By some appropriate method (which is unspecified) partition the given Y vectors into c clusters Ψ_j, $j = 1,2,\ldots,c$, and compute the mean vectors m_j, $j = 1,2,\ldots,c$.

Step 2 (Classification): Select a vector y in Y, and assign it to that cluster whose mean is closest to y. In other words, assign y to Ψ_j if

$$d(y,m_j) = \min_k d(y,m_k).$$

Step 3 (Updating): Update mean vector m_j, $j = 1, 2, \ldots, c$.

Step 4 (Termination): Go to Step 2 unless a complete scan of patterns in Y results in no change in the cluster-mean vectors.

Upon termination of the algorithm, m_j for $j = 1, 2, \ldots, c$ are the templates that are stored in memory. The method described above has been used extensively for the past 20 years [35].

An alternative approach to cluster analysis which is not as computationally intensive, as described below, allows for real-time performance with the vocoder. An advantage for using the vocoder hardware is that the templates can be updated while two-way conversation is in progress. Such an arrangement may be necessary for achieving higher intelligibility at low-bit rates in a multispeaker environment or at a noisy speaker site. Another advantage of the real-time, on-line cluster analysis is that it employs the same front-end (including the microphone, antialiasing filter, and spectral analysis) as used in the actual communications. (We cannot overemphasize the importance of using a matched front end for the training data collection and the actual speech transmission. Some microphones used in the military are far from ideal. In some cases, there is as much as a 20 dB difference between spectral peaks and spectral nulls within the passband.) One plausible algorithm for real-time, on-line clustering analysis is by way of successive dichotomy of each vector from the training set into the following two classes: belonging to an already-established cluster space, or establish a new cluster space with the new sample point as cluster center. Clustering has the following steps:

Step 1: The first vector is treated as the first template, and it is stored in memory.

Step 2: The second incoming vector is compared with the stored template. If the mutual distance measured by the chosen distance criterion is greater than a preset threshold (ideally, it is a just-noticeable distance) then the second vector becomes the second template. Otherwise, no other action is taken.

Step 3: The subsequent incoming vector is compared with every stored template. If the mutual distance between the incoming vector and any one of the stored templates is less than the threshold, no further comparison is needed because it found a cluster to which it belongs. If the mutual distance between the incoming vector and every stored template is greater than the threshold, it becomes a new template.

Step 4: The operation indicated by Step 3 is repeated until the maximum allowable template size is reached. (An exhaustive search of as much as 3840 templates is no longer a problem using state-of-the-art signal processors.)

Both clustering algorithms have been successfully implemented by others in 800-b/s vocoders. We prefer the latter approach for clustering because we are working toward the implementation of a lower bit-rate vocoder (target rate of 300 b/s). We feel some form of automatic template updating is essential to achieve acceptable speech quality at this low rate.

To make the cluster hyperspace smaller for a large size of training data, we experimented with two sets of templates: one for male voices (i.e., pitch frequency of 200 Hz or less), and the other for female voices (i.e., pitch frequency greater than 200 Hz). By using the two sets of templates, the male DRT

score remained virtually unchanged (87.0 as shown in the next section), but the DRT score for a female voice improved by two points. We need more experimentation along this line before we can justify doubling the template memory size.

DRT Scores

A low-bit-rate vocoder, such as our 800-b/s line-spectrum vocoder, does not reproduce speech clearly enough for general use such as the telephone. On the other hand, it is good enough for tactical use where the messages are comparatively more structured than everyday conversations. Some four years ago, we had an experimental real-time 800-b/s vocoder. Although it had a poor DRT score (only 78.3 for three male speakers, "LL," "CH," and "RH"), we were able to communicate in a two-way conversational setup. Now the DRT score for the line-spectrum vocoder operating at the same data rate is 87.0 for the same speakers (Table 11).

Table 11 — Three Male DRT Scores of the 800-b/s Line-Spectrum Vocoder. This table lists the DRT scores for three male speakers ("LL," "CH," and "RH") for our line-spectrum vocoder operating at fixed and synchronous data rate of 800 b/s. For comparison, the DRT score for a 2400-b/s LPC is also shown. Both vocoders use identical sets of unquantized reflection coefficients, voicing decisions, and speech rms values as input data. At the time of writing this report (June 1984), a DRT score of 87 is probably the highest attained by any vocoder operating at the fixed data rate of 800 b/s. The LSF templates do not contain any LSFs generated from these three DRT speakers.

Sound Classification	800 b/s	2400 b/s	Differential
Voicing	89.3	91.7	−2.4
Nasality	89.8	93.7	−3.9
Sustention	79.2	79.7	−0.5
Sibilation	92.4	89.6	+2.8
Graveness	79.2	80.7	−1.5
Compactness	91.9	94.8	−2.9
Total	87.0	88.4	−1.4

The relatively high overall DRT score for our line-spectrum vocoder is a result of the relatively high attribute scores for "graveness" and "sustention." A higher score for "graveness" implies that the speech spectral envelope is well characterized by our filter-parameter quantizer. A higher score for "sustention" implies that the speech parameters rise quickly at abrupt onsets. We believe that no single factor has contributed to the overall enhancement in the DRT score, but rather it is a result of the accumulative but important steps we have made. In general, vocoders over which we can communicate easily have higher DRT scores according to communicability tests conducted at NRL [1]. Hence, it is essential to realize a satisfactory DRT prior to committing any prototype vocoder into production.

4800-b/s Encoder/Decoder

A 4800-b/s voice processor is needed to provide communicators with improved speech quality when compared to the conventional 2400-b/s LPC. To achieve this objective, a 4800-b/s voice processor must be free from the most serious limitations inherent in the 2400-b/s LPC; namely, the speech

waveform being classified into either a periodic or aperiodic waveform. This type of speech classification for narrowband communication has been the tradition since the first vocoder was devised some 50 years ago. It requires both pitch tracking and voicing-state estimation (Fig. 2). Pitch tracking is difficult to accomplish, and it tends to smooth the pitch contour. Natural speech has many pitch irregularities, particularly at voiced onsets and vowel-to-vowel transitions. Without the natural-pitch irregularities, speech tends to sound mechanical. Likewise, voicing-state estimation is even more difficult to determine than pitch tracking, particularly with breath noise and low-frequency, dominant-unvoiced plosives such as /p/.

Recently, however, some progress has been made toward eliminating grotesque pitch and voicing errors. The use of high-speed digital signal processing has made possible more complex arithmetic operations, elaborate logic operations, and delay decisions. Yet, the 2400-b/s LPC still makes occasional pitch and voicing errors. For example, pitch doubling can briefly cause a male voice to sound like a female voice. Voicing error can cause breath noise to be reproduced as a snore and /p/ to approximate the sound of a bilabial fricative. These are some of the more significant reasons why the narrowband LPC has not been universally accepted by the general user. Even experienced communicators who accept distorted CB sounds have reservations about narrowband-LPC sounds.

Thus, we will eliminate the use of both pitch and voicing in our 4800-b/s line-spectrum vocoder as we have done previously for our 9600-b/s multirate processor [36]. This high-rate device has been implemented twice for real-time operation, and it has been extensively tested. In terms of the DRT, it ranks on par with the 16,000-b/s CVSD [36]. In terms of communicability, it is much closer to a 32,000-b/s CVSD than the 2400-b/s LPC (Fig. 1). According to a recent speaker-recognition test, the 9600 b/s scored as high as a 64,000-b/s pulse-code modulator (PCM). We also note that the communicators over this device indicated their satisfaction in regards to the ease and effort needed to communicate. Thus, we are not misguided if we design the 4800-b/s line-spectrum vocoder on the same principle as our previous 9600-b/s processor.

But we have to accomplish a two-to-one data-rate compression in order to reduce 9600 b/s to 4800 b/s. A partial solution comes from the use of LSFs as filter parameters rather than reflection coefficients which were used for the 9600-b/s data rate. Another partial solution also comes from the use of a more coarsely quantized excitation signal than that used in the 9600-b/s processor. Table 12 is an example of a bit allocation for the 4800-b/s line-spectrum vocoder. These figures are justified in the following discussion.

LSF Encoder/Decoder

Since filter parameters are updated once per frame (i.e., 180 speech-sampling-time intervals), relatively few bits are required to encode them. For the 800-b/s line-spectrum vocoder, only 12 bits are used as discussed earlier. It is easy to show that the use of a few extra bits for the filter parameters could improve the output speech quality significantly. On the other hand, the excitation signal is a sample-by-sample parameter. Thus, for nonpitch excitation many bits are required to encode them, and we have allocated as many as 85 bits (Table 12) for the 4800-b/s line-spectrum vocoder. It is also easy to show that using a few extra bits for the excitation signal often does not noticeably enhance the output speech quality. It is difficult to know exactly how many bits should be allocated for filter parameters and the excitation signal because the output speech quality is influenced by both. As an example, both the 2400-b/s and 9600-b/s processors illustrated in Fig. 1 use identical filter parameters, yet performance at 9600 b/s is far superior because it uses more bits for the excitation signal [36]. Our goal is to make the speech-synthesis filter as good as that of the 2400-b/s LPC. The 2400-b/s LPC uses 41 bits to encode 10 reflection coefficients. If we were to encode 10 LSFs instead, we can save 10 bits without hurting the DRT score (Table 5). These LSFs are quantized to have frequency resolution, and the 10 LSFs are selected from the 42 frequencies listed in Table 13. We will use such a frequency quantization rule for encoding the LSFs.

Table 12 — Bit Allocation for Filter Parameters and Excitation Signal. Each parameter is renewed once per frame at a rate of 44.444 Hz. Note that the 2400-b/s LPC uses comparatively few bits to encode the excitation-signal parameters. In contrast, the higher rate devices use nearly four times more bits to encode the excitation signal than the filter parameters.

Parameters	2400-b/s[a] (bits)	9600-b/s[b] (bits)	4800-b/s[c] (bits)
Filter	41	41	21
Excitation Signal	12	172	85
Sync	1	3	2
Total	54	216	108

[a]Government-Standard 2400-b/s LPC.
[b]Navy Multirate Processor [36].
[c]Line-spectrum vocoder.

Table 13 — List of Frequencies with a 6% Frequency Resolution. All frequencies above 400 Hz are quantized at 12 steps per octave (i.e., equitempered chromatic scale), and rounded off to 10 Hz. The two frequencies below 400 Hz do not obey this rule because they occur in murmurs, breath noise, etc. which are not critical elements of normal speech.

Index	Freq. (Hz)	Index	Freq. (Hz)	Index	Freq. (Hz)	Index	Freq. (Hz)	Index	Freq. (Hz)
1	300	3	400	15	800	27	1600	39	3200
2	350	4	420	16	850	28	1700	40	3390
		5	450	17	900	29	1800	41	3590
		6	480	18	950	30	1900	42	3810
		7	500	19	1010	31	2020		
		8	530	20	1070	32	2140		
		9	570	21	1130	33	2260		
		10	600	22	1200	34	2400		
		11	640	23	1270	35	2540		
		12	670	24	1350	36	2690		
		13	710	25	1430	37	2850		
		14	760	26	1510	38	3020		

Although we can encode LSFs independently, we chose to encode center and offset frequencies of LSF pairs, as defined by Eqs. (56a) and (56b) because of the following advantages:

- the highest-offset frequency can be eliminated from encoding because it is least significant, as noted from Table 7;
- fewer bits can represent all the center frequencies because their distributions, as illustrated by Fig. 11, are almost nonoverlapping, particularly for voiced speech;
- the offset frequency of a spectrally sensitive, closely spaced LSF pair is well preserved because the offset frequency is independently quantized with a minimum step of one unit in terms of frequency codes;
- and transmission bit errors affect the frequency response of the synthesis filter in a relalatively small frequency range.

With some clamping of the upper and lower ranges, center and four offset frequencies of LSF pairs may be encoded as indicated in Table 14. The total number of bits to encode filter parameters is only 21.

Table 14 — Encoded Filter Parameters and Their Ranges for 4800-b/s Line-Spectrum Vocoder. The total number of bits used is 21, 20 bits less than that used for the 2400-b/s LPC, but 9 bits more than that used for the 800-b/s line-spectrum vocoder.

Filter Parameters		Frequency Index[a]	No. of Bits
Center Frequency of LSF Pair	1	3, 4, 5, 6, 7, 8, 9, 10, 11 12, 13, 14, 15, 16, 17, 18	4
	2	17, 18, 19, 20, 21, 22, 23, 24	3
	3	25, 26, 27, 28, 29, 30, 31, 32	3
	4	33, 34, 35, 36	2
	5	38, 39, 40, 41	2
Offset Frequency of LSF Pair	1	1, 2, 4, 6	2
	2	1, 2, 3, 4,	2
	3	1, 2, 3, 4,	2
	4	1, 2	1
	5	1 (fixed)	0
			Total ... 21

[a]See Table 13.

Excitation-Signal Encoder/Decoder

The ideal excitation signal for the LPC analysis/synthesis system is the prediction residual because it can produce output speech identical to the input speech in the absence of quantization. Virtually all LPC-based, high-rate voice processors derive their excitation signals from the prediction residual. In principle, these residual encoding techniques are applicable to our 4800-b/s line-spectrum vocoder. Unfortunately, we have only 85 bits to encode the prediction residual, whereas our 9600-b/s processor has twice as many bits. Not all the residual encoding techniques are usable when the number of available bits is only 85. Our residual-encoding technique is based on several levels of scrutiny.

First, we have to decide whether the entire residual samples should be transmitted with coarser quantization, or partial residual samples (typically those occupying below 1 kHz) should be transmitted with finer quantization. If the lowband-residual samples are transmitted, the upper residual samples must be regenerated at the receiver. According to communicability tests conducted at NRL [1], the lowband-residual excited LPC is preferred over the wideband-residual excited LPC because there is less audible-quantization noise at the output.

Once the lowband-residual excitation approach is selected, there are still two possible ways of encoding residual samples: encoding time samples or spectral components. We chose the latter approach because of the following advantages: (a) low-pass filtering and down sampling are not required; (b) low-frequency components below 250 Hz, not essential to speech communications, can be

readily eliminated to save as much as 24 bits (i.e., 6 spectral components at 4 bits each); (c) bit-tradeoff between speech data and overhead data (sync bits, amplitude normalization factor, etc.) is more flexible because a reduction of one-speech-spectral component creates a small-data package of 4 bits to encode overhead data; and (d) upper frequency components may be regenerated by simple spectral replication.

One drawback of this spectral-encoding method is that we need a time-to-frequency transformation of the prediction residual. To obtain spectral components of the prediction residual, we perform the following operations. The 12 trailing-residual samples from the preceding frame are overlapped with the 180 residual samples of the current frame to reduce waveform discontinuity at the frame boundary. Then, the 192 overlapped samples are "trapezoidally windowed" with linear-amplitude weighting over the 12 overlapped samples. The time-to-frequency transform is carried out by the use of a 96-point (FFT). The use of a half-size FFT reduces computations because the input-residual samples are real, and we need spectral components only below 1 kHz. The maximum amplitude spectral component below 1 kHz is transmitted as the amplitude-normalization factor. It is quantized to one of thirty-two 1.75-dB steps covering a dynamic range of 56 dB. Thus 5 out of 86 bits are used as overhead data. The remaining 80 bits are used for encoding 20 spectral components, the 7th through 26th components. Since the frequency separation is (4000/96) = 41.667 Hz, the lowband-residual information covers the frequency range from 250 Hz to 1041.67 Hz.

Each of these 20 spectral components may be encoded in terms of its real and imaginary parts, or in terms of its amplitude and phase spectral components. We note that preservation of phase information is vital to the synthesis of high-quality speech because it defines how each spectral component is phased in reference to the LPC frame which is not pitch synchronous. Thus, encoding the amplitude and phase components is preferred. Although they may be encoded independently, we chose to encode them jointly because of the following advantages: (a) the number of amplitude steps can be traded with the number of phase steps for improved speech quality; (b) an amplitude-dependent phase resolution is feasible (i.e., if the amplitude-spectral component is -15 dB or less with respect to the amplitude-normalization factor, then the corresponding phase component may be quantized more coarsely because we cannot hear it as well as the other components); and (c) we can have more diversified phase angles for more natural sounding speech.

The available 80 bits are equally divided for encoding the 20 complex-spectral components whose amplitudes have been normalized by the maximum amplitude spectral component (i.e., all magnitudes are less than or equal to unity). Thus, encoding each spectral component with 4 bits is equivalent to selecting one of 16 encoding points located within a unit circle. These spectral-encoding points are designed from the probability-density functions of both the residual amplitude and phase spectral components. Since the LPC-analysis frame is not pitch synchronous, the probability-density function of the residual-phase-spectral components is random, and it is uniformly distributed between $-\pi$ and π radians. Thus, a uniform quantizer may be used for phase encoding. The phase resolution is amplitude-dependent, as will be shown. On the other hand, the probability density function of the residual-amplitude-spectral components is bell-shaped as shown in Fig. 20. Thus, the amplitude quantizer will have unequal step sizes.

For a 4-bit quantizer of a complex-spectral component, 2 bits may be assigned for amplitude resolution. But, according to our experimentation, a four-level amplitude quantizer does not leave enough room for adequate phase resolution. We prefer the use of a three-level amplitude quantizer. The design of this quantizer is based on the following amplitude transfer characteristics:

$$\begin{aligned} y(x) &= x_1/2 && \text{if } 0 \leqslant x \leqslant x_1, \\ &= (x_1 + x_2)/2 && \text{if } x_1 < x \leqslant x_2, \\ &= (x_2 + 1)/2 && \text{if } x_2 < x \leqslant 1, \end{aligned} \tag{66}$$

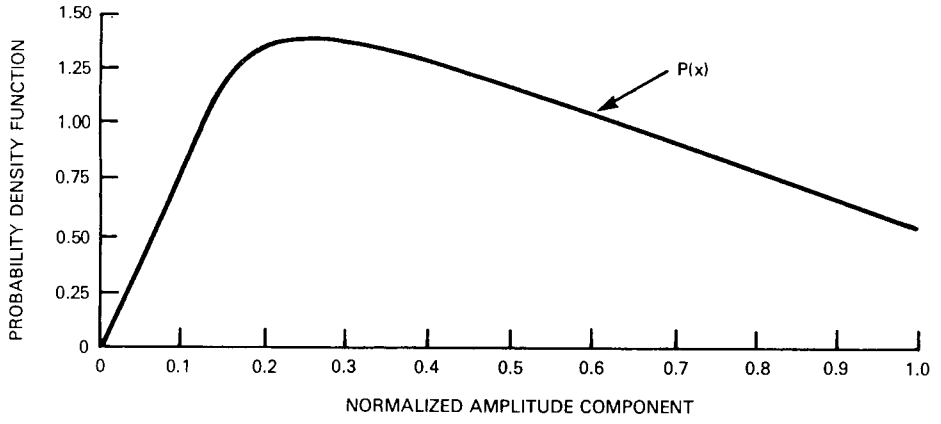


Fig. 20 — Probability density function of residual-amplitude-spectral components (amplitude normalized) from voiced speech. This curve was obtained from 1,600,000 amplitude-spectral components from both male and female voices.

where x is the normalized-input amplitude, $y(x)$ is the output amplitude, and x_1 and x_2 are input-amplitude-break points.

The quantization error is defined as the quantized output amplitude minus the input amplitude:

$$\epsilon(x) = y(x) - x\,. \tag{67}$$

The mean-square value of the quantization error is

$$\bar{\epsilon}^2 = \sum_{x=0}^{x_1} (y(x) - x)^2 p(x) + \sum_{x=x_1}^{x_2} (y(x) - x)^2 p(x) + \sum_{x=x_2}^{1} (y(x) - x)^2\, p(x). \tag{68}$$

The quantizer parameters (x_1 and x_2) which minimize the above error have been computed. The resulting quantizer has the amplitude transfer characteristic:

$$\begin{aligned} y(x) &= 0.172 \quad \text{if } 0.000 \leqslant x \leqslant 0.344, \\ &= 0.500 \quad \text{if } 0.344 < x \leqslant 0.656, \\ &= 0.828 \quad \text{if } 0.656 < x \leqslant 1.000. \end{aligned} \tag{69}$$

Figure 21 shows a constellation of the 16 spectral-encoding points for each complex-spectral component where radii of the three rings are 0.172, 0.500, and 0.828, as already listed above. Phase resolution is amplitude-dependent; from the outermost ring to the innermost ring there are seven, five, and three phases.

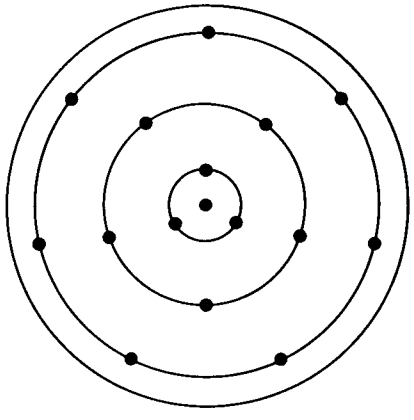


Fig. 21 — Constellation of spectral-encoding points for a 4-bit quantizer

Output Speech Evaluation

Ten years ago a 64,000-b/s nonpitch-excited vocoder was deployed in a limited quantity. Although people found it easy to talk over, its DRT score was actually below that of a 2400-b/s vocoder available at that time. We feel that intelligibility is the most important performance index for a low-bit-rate voice processor because it is often deployed in tactical communications where conversations are generally brief. In some cases, there may be no time to request the message again. Fortunately, our 4800-b/s line-spectrum vocoder scored somewhere between the 2400-b/s LPC and the 9600-b/s multirate processor, as should be (Table 15).

Table 15 — Three Male DRT Scores of 4800-b/s Line-Spectrum Vocoder. As before, the three male speakers are "LL," "CH," and "RH." For comparison, DRT scores of a 2400-b/s LPC and the 9600-b/s Navy multirate processor are also listed.

Sound Classification	2400 b/s	4800 b/s	9600 b/s
Voicing	91.7	94.8	96.3
Nasality	93.7	97.4	99.2
Sustention	79.7	91.1	88.3
Sibilation	89.6	93.2	92.5
Graveness	80.7	82.0	84.4
Compactness	94.8	95.1	97.4
Total	88.4	92.3	93.0

One strength of a nonpitch-excited voice processor is that it performs much better with noisy speech (when the input speech is noisy, the output speech is similarly noisy). In contrast, the 2400-b/s pitch-excited LPC is incapable of making similar noisy speech because the pitch-excitation signal (i.e., pulse train) does not contain sample-by-sample noise. Figure 22 vividly illustrates the difference between the 4800-b/s line-spectrum vocoder output and the 2400-b/s LPC output when the input speech is noisy. The sound difference is even more striking than the visual contrast revealed in Fig. 22.

CONCLUSIONS

Reflection coefficients have been the most often used filter parameters to represent the speech synthesizer in an all-pole-filter configuration. This report presents equivalent filter parameters, called line-spectrum frequencies, which are frequency-domain parameters. Thus, frequency-dependent hearing sensitivities can be incorporated into the quantization process so as to represent crudely something that is not readily discernible to the human ear.

A benefit of using line-spectrum frequencies is that the same level of initial consonant intelligibility is achieved by using 10 bits (approximately 25%) less than that required by reflection coefficients. Furthermore, speech degradation at additional bit savings is gradual. As a result, line-spectrum frequencies may be used for implementing an 800-b/s voice processor which is capable of providing speech intelligibility several points higher than other 800-b/s voice processor heretofore tested. Likewise, line-spectrum frequencies may be used for implementing a 4800-b/s voice processor which is free from the use of both the pitch and voicing decision (i.e., nonpitch-excited narrowband voice processor). Both are welcome additions to narrowband voice processors: the 800-b/s line-spectrum vocoder for transmitting speech over more constricted-information channels, and the 4800-b/s line-spectrum vocoder for achieving a higher communicability than achievable with the 2400-b/s voice processor.

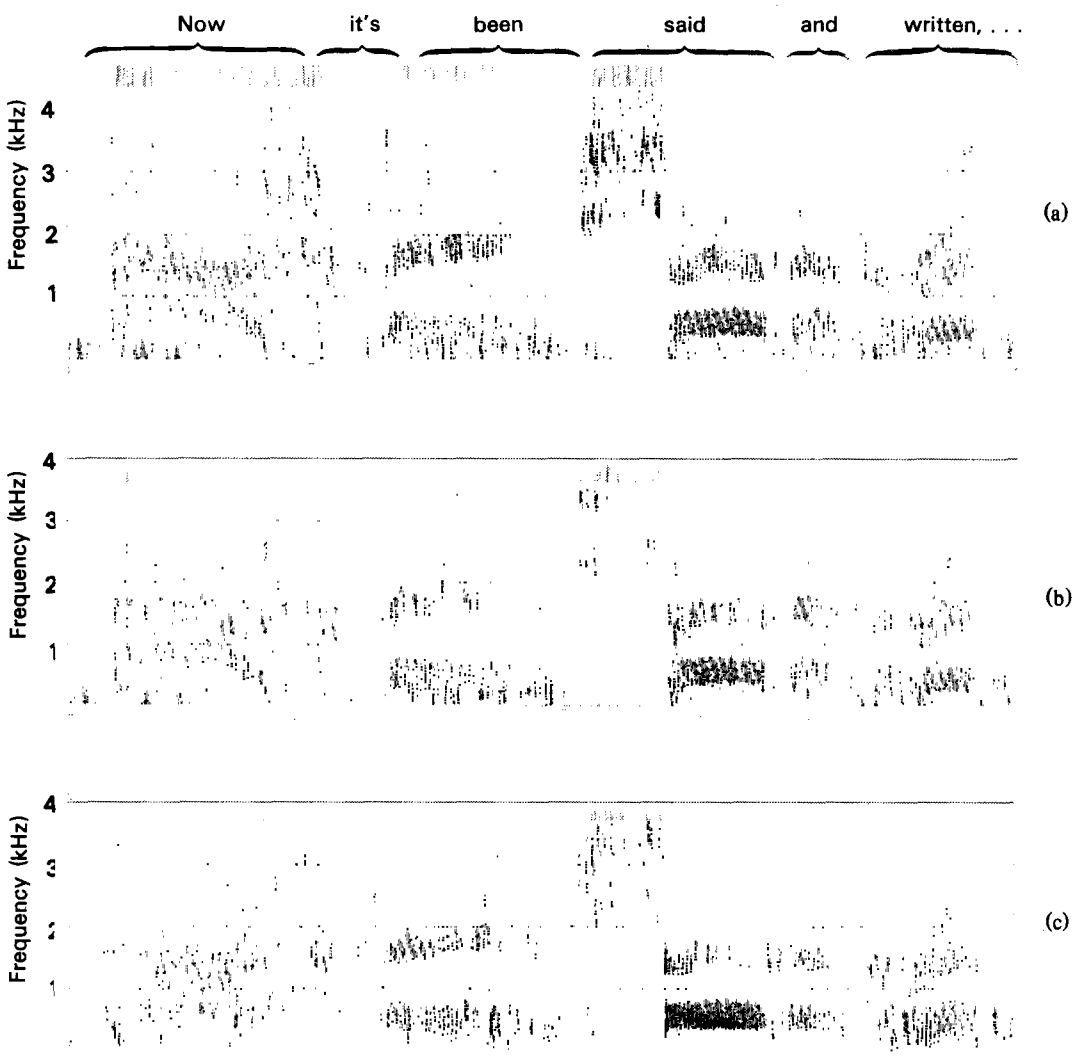


Fig. 22 — Spectrum of noisy input speech, output speech from the 4800-b/s line-spectrum vocoder, and output speech from the 2400-b/s LPC. The input speech was taken from a commercial record in which a newsman is asking a question to the then-President John F. Kennedy in a White House conference room. The output from the 4800-b/s line-spectrum vocoder is a closer replica of the original than the output of the 2400-b/s LPC. Note the absence of noise in the voiced segment of the 2400-b/s LPC. (a) Original speech; (b) output speech from 4800-b/s line-spectrum vocoder; (c) output speech from 2400-b/s LPC.

ACKNOWLEDGMENTS

This work is funded by Office of Naval Research and NAVELEX. The authors thank Drs. J. Davis and B. Wald of NRL, and R. Martin and R. Allen of NAVELEX for their support.

REFERENCES

1. A. Schmidt-Nielsen and S.S. Everett, "A Conversational Test for Comparing Voice Systems Using Working Two-Way Communication Links," *IEEE Trans. Acoustics, Speech and Signal Proc.* **ASSP-30**, 853-863 (1982).

2. H. Dudley, "Signal Transmission," U.S. Patent 2, 151, 091 (1939).

3. D.B. Paul, "The Spectral Envelope Estimation Vocoder," *IEEE Trans. on Acoustics, Speech and Signal Proc.* **ASSP-29**(4), 786-794 (1981).

4. J.L. Flanagan, C.H. Coker, L.R. Rabiner, R.W. Shafer, and N. Umeda, "Synthetic Voices for Computers," *IEEE Spectrum*, 20-41 (1970).

5. B. Gold and L.R. Rabiner, "Analysis of Digital and Analog Formant Synthesizer," *IEEE Trans. Audio and Electroacoust* **AU-16**(1), 81-94 (1968).

6. J.N. Holmes, "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer," *IEEE Trans. on Audio and Electroacoust.* **AU-21**, 298-305 (1973).

7. B.S. Atal and S.L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.* **50**(2) (Part 2), 637-655 (1971).

8. E.I. Jury, *Theory and Application of the z-Transform Method* (John Wiley & Sons, Inc., New York, 1964).

9. F. Itakura, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals," *J. Acoust. Soc. Am.* **57**, Supplement No. 1, S35 (1975).

10. T. Kobayashi, "Speech Synthesis and Speech Recognition LSI," *J. Acoustics Soc. Japan* **39**(11), 744-749 (1983) (in Japanese).

11. F. K. Soong and B. Juang, "Line-Spectrum Pair (LSP) and Speech Data Compression," *IEEE International Conference on Acoustics, Speech and Signal Processing, IC***ASSP-83**, CH1945-5, 1.10.1-1.10.4 (1983).

12. H. Wakita, "Linear Prediction Voice Synthesizers: line-spectrum pairs (LSP) is the newest of several techniques," *Speech Technology*, 17-22, fall 1981.

13. A. Papoulis, *The Fourier Integral and Its Applications* (McGraw-Hill Book Co., Inc., New York, 1982).

14. K.J. Astrom, *Introduction to Stochastic Control Theory* (Academic Press, Inc., New York, 1970).

15. S.A. Tretter, *Introduction to Discrete-Time Signal Processing* (John Wiley & Sons, Inc., New York, 1976).

16. J. Makhoul, "Linear Prediction; A Tutorial Review," *Proc. IEEE* **63**(4), 561-580 (1975).

17. J.W. Fussell, M.D. Cowing, P.W. Boudra, Jr., and B.M. Abzug, "Providing Channel Error Protection for a 2400-bps Linear Predictive Coded Voice System," *IEEE: IC***ASSP-78**, CH 1285-6, 462-465 (1978).

18. M.R. Schroeder, B.S. Atal, and J.L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," *J. Accoust. Soc. Am.* **66**(6), 1647-1652 (1979).

19. B.S. Atal, "Predictive Coding of Speech Signals and Subjective Error Criteria," *IEEE Trans. on Acoustics, Speech, and Signal Proc.* **ASSP-27**(3), 247-254 (1979).

20. P. Ladefoged, *Elements of Acoustic Phonetics* (The University of Chicago Press, Chicago and London, 1974).

21. H.R. Schiffman, *Sensation and Perception: An Integrated Approach* (John Wiley & Sons, Inc., New York, 1976).

22. J.L. Flanagan, *Speech Analysis, Synthesis and Perception* (Springer-Verlag, New York, 1972).

23. R.T. Gagnon, "VOTRAX Real Time Hardware for Phoneme Synthesis of Speech," *IEEE IC***ASSP-78**, CH 1285-6, 175-178 (1978).

24. R. Wiggins and L. Brantingham, "Three-Chip System Synthesizes Human Speech," *Electronics,* 109-116 (1978).

25. R.K. Potter, G.A. Knopp, and H.G. Kopp, *Visible Speech* (Dover Publications, Inc., Mineola, N.Y., 1966).

26. B. Gold and J. Tierney, "Vocoder Analysis Based on Properties of Human Auditory System," Technical Report 670, Lincoln Laboratory, MIT, Dec. 1983.

27. P. Strevens, "Spectra of Fricative Noise in Human Speech," *Language and Speech* **3**, 202-218 (1960).

28. V.A. Mann and B. Repp, "Influence of the Vocalic Context on Perception of the [S]-[s] Distinction," *Perception and Psychophysics* **28**, 213-228 (1980).

29. L.J. Fransen, "2400- to 800-b/s LPC Rate Converter," NRL Report 8716, June 1983.

30. C.P. Smith "Perception of Vocoder Speech Processed by Pattern Matching," *JASA* **46**(6) (Part 2) 1562-1571 (1969).

31. B. Gold, "Experiments with a Pattern-Matching Channel Vocoder," *IEEE IC***ASSP-81**, CH 1610-5, 32-34 (1981).

32. D.B. Paul, "An 800-bps Adaptive Vector Quantization Vocoder Using a Perceptual Distance Measure," *IEEE IC***ASSP-83**, CH 1841-6, 73-76 (1983).

33. D.Y. Wong, B. Juang, and A.H. Gray, Jr., "An 800-bit/s Vector Quantization LPC Vocoder," **ASSP-30**(5) 770-780 (1982).

34. P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach* (Prentice Hall International, Englewood, N.J., 1982).

35. G.H. Ball and D.J. Hall, "ISODATA. An Iterative Method of Multivariate Data Analysis and Pattern Classification," *IEEE Int. Commun. Conf.,* 116-117 (1966).

36. G.S. Kang, L.J. Fransen, and E.L. Kline, "Medium Band Speech Processor with Baseband Residual Spectrum Encoding," *IEEE IC***ASSP-81**, 820-823 (1981).

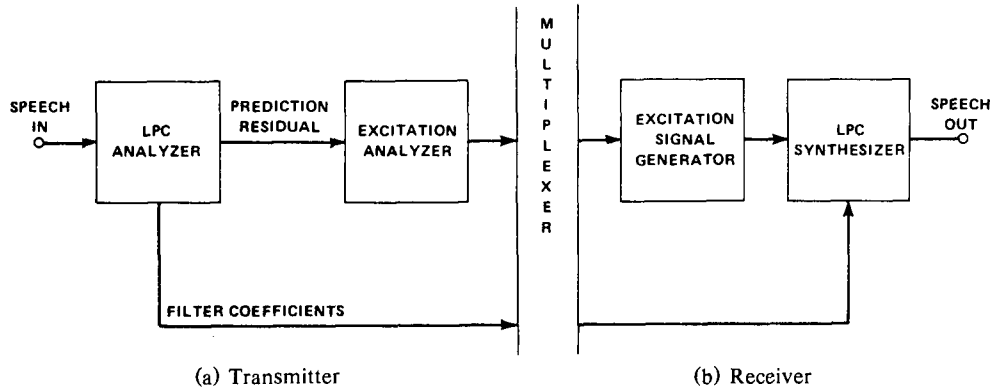
Appendix

SUMMARY OF LPC ANALYSIS AND SYNTHESIS

The basic equations of LPC analysis and synthesis are summarized in this appendix. These equations are well known in the speech processing field, but they may be helpful to those getting acquainted to this area.

OVERVIEW OF LPC ANALYSIS/SYNTHESIS

The LPC analysis decomposes a given speech waveform into two component waveforms (Fig. A1). One waveform is a set of slowly time-varying components (i.e., prediction coefficients or *filter coefficients*) which represent the resonance characteristics of the vocal tract. The other waveform is a wideband signal, called the prediction residual, which is the difference between the actual and the predicted speech samples. The prediction residual is an ideal excitation signal for the speech synthesizer because it produces a synthesis filter output nearly identical to the input speech. For a speech transmission rate of 2400 b/s, the prediction residual is modeled as one of two rudimentary signals: a pulse train (or repetitive broadband signal) for voiced sounds and random noise for unvoiced sounds. In essence, the prediction residual is characterized by three *excitation parameters*: pitch period, voicing decision, and amplitude information. The filter coefficients and excitation parameters are updated periodically (every 22.5 milliseconds (ms)).



(a) Transmitter (b) Receiver

Fig. A1 — Block diagram of 2400-b/s LPC

BLOCK-FORM LPC ANALYSIS AND SYNTHESIS

In linear predictive analysis, a speech sample is represented as a linear combination of past samples. Thus,

$$x_i = \sum_{j=1}^{n} \alpha_{j|n}\, x_{i-j} + \epsilon_i \qquad i = 0,1,2,\ldots, m, \tag{A1}$$

where $\alpha_{j|n}$ is the j th prediction coefficient of the n th order predictor, and ϵ_i is the i th prediction residual.

In matrix notation, Eq. (1) can be written as

$$X = H\,\alpha + \epsilon. \tag{A2}$$

On the basis of unbiased estimation (i.e., the covariance matrix of ϵ is an identity matrix), the prediction coefficients are obtained from

$$(H^T H)\alpha = H^T X. \tag{A3}$$

Matrix $H^T H$ may be transformed to LDL^T through Cholesky decomposition, where D is a diagonal matrix, and L is a lower triangular matrix. The solution for the prediction coefficients is carried out in two steps:

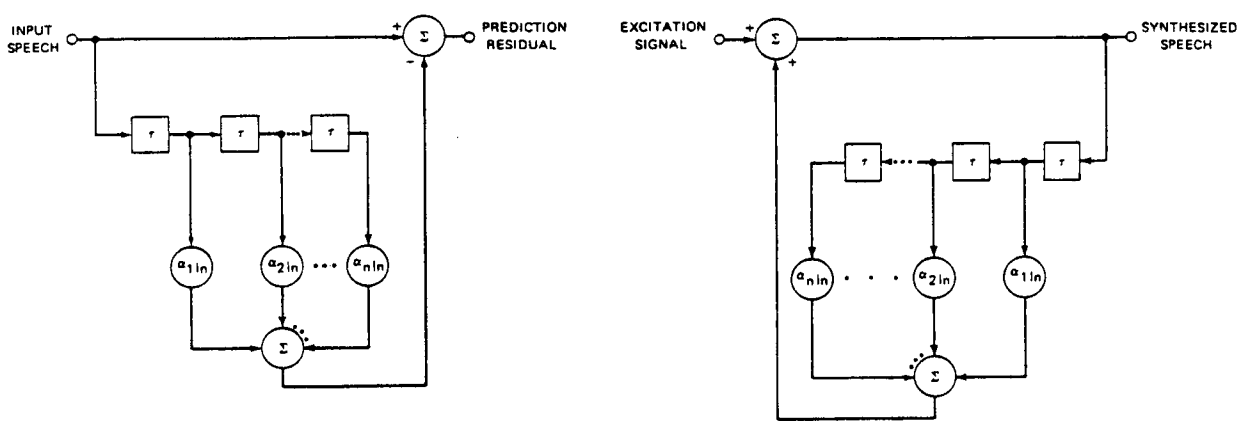
$$k = L^{-1}(H^T X) \tag{A4}$$

and

$$\alpha = (DL^T)^{-1}k. \tag{A5}$$

The quantity k in Eq. (A4) is a set of reflection coefficients, and they are linearly and uniquely related to a set of prediction coefficients by Eq. (A5). All current LPCs transmit reflection coefficients in lieu of prediction coefficients. As mentioned earlier, the excitation signal is parametized by three parameters: pitch period, voicing decision, and amplitude information. Optionally, the amplitude information can be based on speech power rather than residual power.

The synthesizer regenerates speech based on Eq. (1) where ϵ_i is replaced by the artificial excitation signal. The block-form method uses Eqs. (A1) through (A5) to generate an LPC analysis/synthesis speech processor (Fig. A2).



(a) Analysis filter

(b) Synthesis filter

Fig. A2 — Analysis and synthesis filters with prediction coefficients as weights

FLOW-FORM LPC ANALYSIS AND SYNTHESIS

In flow-form LPC analysis, a speech sample is expressed directly in terms of reflection coefficients denoted by k_i. Thus,

$$x_i = \sum_{j=1}^{n} k_j\, y_{i-j} + \epsilon_i \quad i = 0,1,2,\ldots m, \tag{A6}$$

where $\{y\}$ is an orthogonal sequence derived from $\{x\}$.

In terms of inner vector notation, these two sequences are related by

$$(\alpha_{j|n},x_j) = (k_j,y_j) \qquad j = 1, 2, \ldots, n. \tag{A7}$$

By the Gram-Schmidt orthogonalization process, y_j in terms of x_j for $j = 1, 2, \ldots, n$, are:

$$y_1 = x_1 \tag{A8}$$

$$y_2 = x_2 - \frac{(x_2,y_1)}{(y_1,y_1)}\, y_1 \tag{A9}$$

.

.

.

$$y_n = x_n - \sum_{j=1}^{n-1} \frac{(x_n,y_j)}{(y_k,y_j)} y_j. \tag{A10}$$

Equations (A7) through (A10) provide an iterative solution for the reflection coefficients. To find this solution $\alpha_{j|n}$ is expressed in terms of k_j for $j = 1, 2, \ldots, n$. Substituting Eqs. (A8) through (A10) into Eq. (A7) for each value of j, the following expressions may be established:

$$[\alpha_{1|1}] = [1][k_1],$$

$$\begin{bmatrix}\alpha_{1|2}\\ \alpha_{2|2}\end{bmatrix} = \begin{bmatrix}1 & -\alpha_{1|1}\\ 0 & 1\end{bmatrix}\begin{bmatrix}k_1\\ k_2\end{bmatrix},$$

$$\begin{bmatrix}\alpha_{1|3}\\ \alpha_{2|3}\\ \alpha_{3|3}\end{bmatrix} = \begin{bmatrix}1 & -\alpha_{1|1} & -\alpha_{2|2}\\ 0 & 1 & -\alpha_{1|2}\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}k_1\\ k_2\\ k_3\end{bmatrix},$$

and so on. The above expressions can be put into one compact recursive expression:

$$\alpha_{j|n+1} = \alpha_{j|n} - k_{n+1}\, \alpha_{n+1-j|n} \quad j = 1, 2, \ldots, n. \tag{A11}$$

The transfer function of the nth order predictor, denoted by $A_n(z)$, is

$$A_n(z) = -\sum_{j=0}^{n} \alpha_{j|n}\, z^{-j} \qquad \alpha_{0|1} = -1. \tag{A12}$$

Substituting Eq. (A11) into Eq. (A12) gives $A_{n+1}(z)$ in terms of $A_n(z)$. Thus,

$$A_{n+1}(z) = A_n(z) - k_{n+1} z^{-n-1} A_n(z^{-1}) \tag{A13}$$

$$= A_n(z) - k_{n+1} B_n(z), \tag{A14}$$

where

$$B_n(z) = z^{-n-1} A_n(z^{-1}).$$

Likewise, $B_{n+1}(z)$ may be expressed as

$$B_{n+1}(z) = z^{-1}\,[B_n(z) - k_{n+1} A_n(z)]. \tag{A15}$$

Equations (A13) and (A15) define the network structure of the flow-form LPC analysis filter, as shown in Fig. A2(a). Each reflection coefficient is estimated directly at each filter section.

The synthesis filter is an inverse filter of the analysis filter. From Eqs. (A13) and (A15), the synthesis filter transfer functions become

$$A_n(z) = A_{n+1}(z) + k_{n+1}B_n(z) \tag{A16}$$

$$z^{-1}B_n(z) = B_{n+1}(z) + k_{n+1}z^{-1}A_n(z). \tag{A17}$$

The network structure of the synthesis filter is shown in Fig. A2(b).

Each reflection coefficient is computed independently at each analysis filter section. The reflection coefficient is chosen to minimize the filter output in the mean-square sense. The outputs of the nth filter section of its input are

$$a_{n+1,i} = a_{n,1} - k_n b_{n,i-1} \tag{A18}$$

and

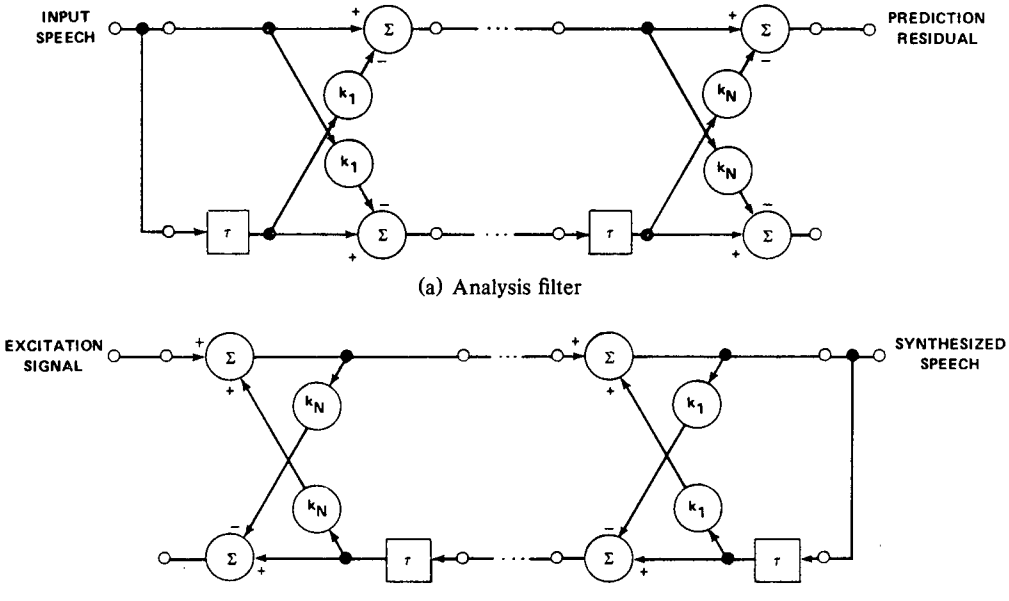
$$b_{n+1,i} = b_{n,i-1} - k_n a_{n,i}, \tag{A19}$$

where $a_{n,i}$ is the upper branch residual (forward residual) and $b_{n,i-1}$ is the lower branch residual (backward residual) as indicated in Fig. A3(a). The reflection coefficient k_n is chosen to minimize the following quantity:

$$I = E[a_{n,i} - k_n b_{n,i-1}]^2 + E[b_{n,i-1} - k_n a_{n,i}]^2, \tag{A20}$$

where $E[\cdot]$ is a low-pass filtering operation. By differentiating Eq. (A20) with respect to k_n, and setting the resulting expression to zero, k_n is:

$$k_n = \frac{E[a_{n,i}b_{n,i-1}]}{\frac{1}{2}\{E[a_{n,i}^2] + E[b_{n,i-1}^2]\}}. \tag{A21}$$



(a) Analysis filter

(b) Synthesis filter

Fig. A3 — Analysis and synthesis filters with reflection coefficients as weights

Equation (A11) converts a set of reflection coefficients to a set of prediction coefficients. Conversely, a set of prediction coefficients may be derived from a set of reflection coefficients by use of Eq. (A11). Thus,

$$\alpha_{j|n+1} = \alpha_{j|n} - k_{n+1}\alpha_{n+1-j|n} \qquad j = 1,2,\ldots,n. \tag{A11}$$

Letting j be replaced by $n + 1 - j$ in Eq. (A11) gives

$$\alpha_{n+1-j|n+1} = \alpha_{n+1-j|n} - k_{n+1}\alpha_{j|n}. \tag{A22}$$

In Eqs. (A11) and (A22), $\alpha_{n+1-j|n}$ and $\alpha_{j|n}$ are the only two unknowns. Thus, solving for $\alpha_{j|n}$, or alternatively $\alpha_{n+1-j|n}$, gives

$$\alpha_{j|n} = \frac{\alpha_{j|n+1} + k_{n+1}\alpha_{n+1-j|n+1}}{1 - k_{n+1}^2}, \tag{A23}$$

where $j = 1,2,3,\ldots, n$ and $k_{n+1} = \alpha_{n+1|n+1}$. Equation (A23) converts a set of reflection coefficients to a set of prediction coefficients.